Epigenetics in Gene Expression and Development

Barkas, Nikolaos

Awarding institution:
King's College London

Epigenetics in Gene Expression and Development

Submitted by

Nikolaos Barkas

June 2015

To King's College London for the Degree of

Doctor of Philosophy

Department of Medical and Molecular Genetics,

King's College London, School of Medicine,

London, SE1 9RT, U.K.

The work submitted in this thesis is my own

Nikolaos Barkas

Abstract

Epigenetic processes are known to play an important role in the regulation of embryonic development and gene expression. Here we utilise next-generation sequencing and bioinformatics methodologies to investigate the role of epigenetics in two different systems, heart and brain. In heart, the endocardium is a distinct understudied epithelial population of cells that is involved in directing morphogenesis of the myocardium, valve leaflets and trabeculae. We generate whole-genome bisulphite sequencing data for the endocardium and endothelium and compare these data to transcriptomic profiles of these cells. We identify a plethora of differentially expressed genes and differentially methylated genomic regions. Through motif analysis we identify the ETS family of transcriptional activators as likely to play a role in the development of the endocardium.

In brain, we investigate the role of the CTCF and cohesin DNA binding factors in imprinted gene expression by performing high depth allele-specific ChIP-seq for these two factors. We develop a novel bioinformatics approach for performing allele-specific mapping of next-generation sequencing reads and we compare our results with existing data for mouse liver and embryonic stem cells. We note that embryonic stem cells have fewer unique CTCF binding sites consistent with their undifferentiated profile. We examine CTCF and cohesin binding in the vicinity of imprinted loci and note that CTCF and/or cohesin bind to a subset of imprinted regions, suggesting a heterogeneous mechanism for imprinting.

Collectively, our studies examine the role of epigenetics and their interplay with transcription in two distinct systems and identify a variable role for these processes in gene expression and development.

Acknowledgements

First and foremost, I would like to thank my supervisor Professor Rebecca Oakey for guidance and support in the course of this project. It is difficult to overstate her contribution in motivating, guiding and providing help when needed.

My appreciation also goes to the people that taught and helped me in the course of the past four years: Dr Mike Cowley, Dr Adam Prickett and Dr Reiner Schulz.

I would like to thank Dr Benjamin Lehne and Dr Venu Pullabhatla for their guidance with next-generation data processing; Dr Efterpi Papouli and Dr Carolina Gemma for their help with next-generation sequencing technologies and whole-genome bisulphite sequencing protocol respectively. Thanks also goes to Mr Samuele Maria Amante and Dr Sabrina Böhm for assisting with various aspects of laboratory work.

I would also like to thank Professor Scott Baldwin and the members of his laboratory for an exciting and engaging collaboration. Special thanks goes to Mr Kevin Tompkins and Dr Chris Harmenlink for their contributions that formed the basis of much of this work.

None of this would have ever been possible without the long-term moral and financial support of Ms Nassia Kontouzoglou, who encouraged me and provided me with the opportunity to study. Thanks also goes to Ms Sevgi Kozakli for her moral support through my graduate studies.

Contents

List of Figures

List of Tables

Abbreviations

H_0 Null Hypothesis

castaneus *Mus musculus castaneus*

Ab Antibody

AVC Atrioventricular canal

BER Base excision repair

bp basepair

BS-seq Bisulphite conversion coupled with next-generation sequencing

BxC C57Bl6 x *castaneus*

C Cytosine

caC 5-carboxylcytosine

cDNA copy DNA

CGI CpG island

CHD Congenital Heart Defects

ChIP Chromatin Immuno-Precipitation

ChIP-seq Chromatin Immuno-precipitation coupled with Next-generation Sequencing

CpG Cytosine-phosphate-Guanine

CTCF CCCTC-binding factor

CTD C-terminal Domain

CxB *castaneus* x C57Bl6

DBD DNA binding domains

DM Differentiation Media

DMR Differentially Methylated Regions

Dnmt DNA methyl-transferase

dsDNA double stranded DNA

EB Embryoid Body

ECM Extracellular Matrix

EMT Endothelial to Mesenchymal Transformation

ES Embryonic Stem

fC 5-formylcytosine

FDR False Discovery Ratio

FPKM Fragments per Kilobase per Million reads

G Guanine

GO Gene Ontology

H3K27me3 Histone 3 Lysine 27 tri-methylation

H3K36me3 Histone 3 Lysine 36 trimethylation

H3K4me3 Histone 3 Lysine 4 tri-methylation

HAT Histone Acetyl-transferase

HDAC Histone Deacetylase

hmC 5-hydroxymethylcytosine

ICR Imprinting Control Region

K Lysine

LV Left Ventricle

mC 5-methylcytosine

MEF Mouse Embryonic Fibroblast

MHC Myosin Heavy Chain

MICP Multipotent Cardiovascular Progenitor

NFATc1 Nuclear Factor of Activated T-cells 1

ng-seq Next-generation sequencing

NO Nitric Oxide

OFT Outflow Tract

P21 Post-natal day 21

PBS Phosphate Buffered Saline

PE Paired End

PRC2 Polycomb Repressive Complex 2

qPCR quantitative PCR

RNApolII RNA polymerase II

rRNA ribosomal RNA

RT Room Temperature

RV Right Ventricle

SNP Single Nucleotide Polymorphism

T Thymine

TET Ten-eleven translocation

TF Transcription Factor

TSS Transcription Start Site

U Uracil

UPD Uniparental Disomy

WGBS Whole-genome bisulphite sequencing

Chapter 1

Introduction

1.1 Epigenetics

1.1.1 Definition

The term epigenetics was first coined by Conrad Waddington in 1942 as *"the branch of biology which studies the causal interactions between genes and their products, which bring the phenotype into being"* [Goldberg et al., 2007]. Waddington coined this term as a counterpart to the term 'Phenogenetics', which referred to the descriptive study of genetic perturbations [Waddington, 2012]. This original definition of epigenetics only vaguely corresponds to the modern ideas about what constitutes epigenetic phenomena. Instead, the term mostly refers to what is now termed gene regulation and the study of protein function to give rise to the visible phenotype. The original meaning of the term is now largely encapsulated in the field of systems biology. Further to coining the term epigenetics, Waddington introduced the concept of 'epigenetic landscape' [Slack, 2002]. The term refers to an imaginary sloped surface with bifurcations on which a ball rolls. At each bifurcation the ball is pictured to make a binary decision. The surface was originally meant to represent states of a multi-dimentional space of cellular metabolism and the ball a cell during development making successive fate decisions at each bifurcation of the landscape.

The definition of the term epigenetics has since evolved to refer specifically to information in the cell nucleus beyond the information in the primary DNA sequence, with some sources imposing a heritability criterion [Berger et al., 2009]. As such, the term epigenetics now refers to phenomena that include DNA methylation, histone modifications, the deposition of rare histone variants on chromatin and regulation by non-coding RNAs.

The idea of the epigenetic landscape has accordingly changed and it is no longer used to refer to states in the space of cellular metabolism, but rather states in terms of DNA methylation and histone modifications.

1.1.2 Epigenetics in Gene Expression Regulation and Cell Fate Specification

A large body of evidence exists to suggest that epigenetic information correlates with and may be able to influence gene expression regulation and cell fate specification, but the mechanisms of this mode regulation are not well defined, with a few notable exceptions such as CpG methylation in the context of CpG islands (see following Section) [Jones, 2012] .

For an epigenetic mark or process to be involved in specification of cell fate, there is a fundamental requirement that the modification in question can propagate across cell division. It is well established that DNA methylation is conserved across cell division (Section 1.1.3). It is also hypothesised that at least some histone modifications can be inherited across cell divisions, although the details of histone modification preservation are not well established despite models for transmission of some histone marks being proposed and supported by experimental evidence [Bannister and Kouzarides, 2011]. Other epigenetic factors such as non-coding RNAs are less likely to be preserved across cell divisions and no general mechanism for their preservation is known.

More recent evidence suggests that epigenetic marks may also be involved in, and provide a mechanistic explanation for, transgenerational inheritance effects, by escaping erasure during germ line development that affects the majority of the genome [Seisenberger et al., 2012]. Findings in this area remain controversial.

1.1.3 Methylation and Hydroxymethylation of DNA

DNA methylation refers to the addition of a methyl group to the fifth carbon position of Cytosine (C) (see Figure 1.1). The C methylation reaction is catalysed by one of the three DNA methyl-transferase (Dnmt) enzymes: Dnmt1, Dnmt3a and Dnmt3b[1].

Dnmt1 was the first Dnmt described [Bestor et al., 1988] and is primarily regarded to be a maintenance methyl-transferase. It exhibits auto-inhibition for *de novo* methylation and is specifically targeted to replication forks by UHRF1 [Sharif et al., 2007] [Bostick

[1]Dnmt2 catalyses the methylation of aspartic acid tRNA [Goll et al., 2006] and will not be discussed here.

et al., 2007] where it methylates hemi-methylated DNA. Dnmt3a and Dnmt3b are *de novo* methyl-transferases and are responsible for methylation of previously unmethylated DNA during germ cell development [Denis et al., 2011]. The mechanism of *de novo* methylation targeting is likely to involve some level of sequence specificity, guidance by Dnmt3l (a protein highly homologous to other Dnmts but with no catalytic activity) and RNA-directed DNA methylation [Denis et al., 2011].

The mechanism of removal of DNA methylation is not fully established but evidence exists to support that it can be removed either passively (by non-renewal) during cell division or actively. Evidence for passive demethylation exists in the the maternal genome of the zygote [Rougier et al., 1998]. Active inducible demethylation has been demonstrated in non-dividing neurons at the promoters of FGF1 and BDNF in the time scale of minutes [Martinowich et al., 2003] and in the male pronucleus [Wossidlo et al., 2011]. Recent evidence also suggest active demethylation in the maternal zygotic genome [Guo et al., 2014].

The ten-eleven translocation (TET) family of enzymes has been implicated in the removal of DNA methylation. TET enzymes are known to catalyse the oxidation of 5-methylcytosine(mC) to 5-hydroxymethylcytosine (hmC), as well as 5-formylcytosine (fC) and 5-carboxylcytosine [Ito et al., 2011]. hmC has been shown to be demethylated by the action of AID and APOBEC deaminase enzymes and the action of the base excision repair (BER) DNA glycosylation pathway [Guo et al., 2011c]. fC and caC can also be removed by the BER DNA glycosylation [Kohli and Zhang, 2013].

hmC has been found to be particularly abundant in the brain and the embryonic stem cells but is also present in other tissues such as lung and muscle [Bhutani et al., 2011]. Unlike other tissues DNA methylation state in embryonic stem cells has recently been shown to be the result of a dynamic equilibrium between methylation and demethylation [Shipony et al., 2014] and this finding is consistent with the intermediate role of hmC and its high abundance in ESCs. The role of hmC in brain is less clear, although it has been implicated in neuronal plasticity [Guo et al., 2011b].

The function of DNA methylation is not clear in most genomic contexts [Jones, 2012]. However, the function of CpG methylation in CpG island promoters (see below) is one of the best understood examples of epigenetic regulation of gene expression. In this context methylation is associated with long-term silenced genes and a chromatin state not permissive to transcription, such as genes on the inactive X chromosome. Furthermore,

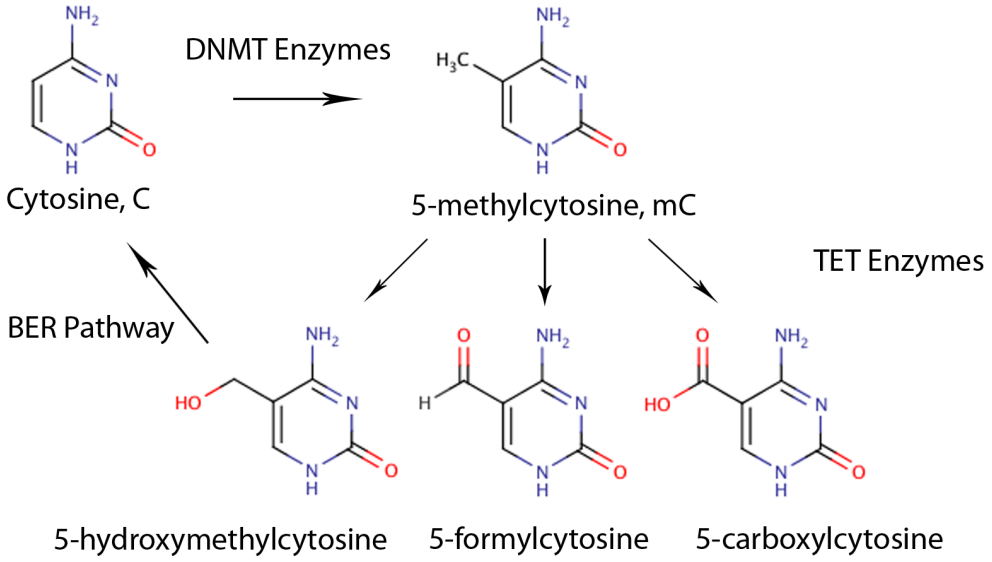


Figure 1.1: Overview of known cytosine modifications catalysed by the DNMT and TET enzymes families. The DNMT family of enzymes catalyses methylation of cytosine at the 5 carbon position. Enzymes in the TET family catalyse further modifications including 5-hydroxymethylcytosine. 5-hydroxymethylcytosine can be converted to cytosine, through the BER pathway.

DNA methylation is known to be important for silencing of transposable elements [Smith and Meissner, 2013].

The function of DNA methylation in non-CpG island promoters and transcribed regions is not understood in vertebrates [Lister et al., 2009]. More recently non-CpG methylation has been shown to be mostly confined to embryonic stem cells [Arand et al., 2012].

CpG Islands

In vertebrates, the majority of cytosine methylation occurs primarily in the context of C followed by a Guanine(G) base (CpG context) [Deaton and Bird, 2011]. This arrangement is symmetrical, when one considers both DNA strands, and allows for preservation of the methylation state upon DNA replication by copying information for the template strand to the newly synthesised strand.

Mutation of cytosines occurs by deamination. Whereas non-methylated Cs can be repaired by the cellular machinery, mCs cannot be reliably repaired and tend to be converted to Ts and lost [Bird, 1980]. As a result, the frequently methylated CpG dinucleotide is significantly underrepresented in mammalian genomes. CpGs in mammals are

only encountered at the expected frequency, and thus overrepresented in comparison with the rest of the genome, in regions termed CpG islands (CGIs), where they are usually unmethylated [Deaton and Bird, 2011].

The majority of CGIs are found in promoter regions of genes, and the presence of sequence elements results in an open, nucleosome depleted chromatin conformation and favours transcriptional activation. CGIs are also frequently associated with histone H3 trimethylation at lysine 4 (H3K4me3). High methylation levels of CGIs are correlated with long-term transcriptional repression [Deaton and Bird, 2011] [Jones, 2012].

CGIs are also found in intragenic regions but the role of intragenic CGIs is less clear, although roles in alternative promoter usage [Jones, 2012] and alternative RNA processing have been proposed [Maunakea et al., 2013]. Extensive variation in methylation surrounding CGIs has been reported and these regions have been termed CGI shores. CGI shores have been implicated in gene regulation in cancer and may also be relevant in normal development [Hansen et al., 2011].

1.1.4 Histone Modifications

DNA in the eukaryotic nucleus is wrapped around histone octamers composed of two copies of each of H2A, H2B, H3 and H4 histone polypeptides. Approximately 147 base-pairs (bp) are wrapped around each octamer with a small variable linker of approximately 80 bps between them. The solution of the histone octamer structure in 1997 showed that the DNA is wrapped around a central core, with N-terminal chains protruding out of the octamer [Luger et al., 1997].

Histones are known to be heavily and dynamically post-translationally modified in the nucleus [Bannister and Kouzarides, 2011]. The majority of the modifications occur on residues on the N-terminal chains, with only a small proportion on the central core residues. A very large repertoire of histone modifications has been described: acetylation, phosphorylation, methylation, deimination, beta-N-acetylglucosamine addition, ADP ribosylation, ubiquitination and sumoylation, proline isomerisation and even histone tail clipping [Bannister and Kouzarides, 2011]. The action of the majority of these modifications is unknown, with most of our understanding focused on a small subset of modifications.

Some of the known histone modifications are well understood and correlate with gene activity. In particular, some forms of lysine (K) methylation are present in well defined

chromatin environments. Histone 3 Lysine 27 tri-methylation (H3K27me3) correlates strongly with facultative heterochromatin and gene expression repression during development but not with constitutive heterochromatin. Histone 3 Lysine 4 tri-methylation (H3K4me3) is encountered in the transcription start site (TSS) of actively expressed genes [Bannister and Kouzarides, 2011]. H3K4me3 is established co-transcriptionally by the action of scSet1 methyl-transferase, which specifically binds to the elongating form of the RNA polymerase II (RNApolII) via the Serine 5 phosphorylation of the C-terminal domain (CTD). Histone 3 Lysine 36 trimethylation (H3K36me3) is also encountered in active genes, but in the bodies rather than the TSS. Furthermore, Histone 3 Lysine 4 mono-methylation is associated with active enhancers and H3K27ac has been proposed to distinguish active from poised enhancer elements [Creyghton et al., 2010].

In contrast to DNA methylation, there is no well established mechanism for the maintenance of the chromatin state across cell divisions, although chromatin state appears to be preserved. Furthermore, it is not clear if all histone modifications are preserved to the same extent. Some mechanisms have been proposed for particular modifications. For example H3K27me3 may be preserved by recruitment of the Polycomb Repressive Complex 2 (PRC2) at the replication fork by the histone mark leading to deposition of the same histone mark at the newly synthesised DNA [Hansen et al., 2008].

1.2 Gene Regulation and Transcription Factor Binding

Gene expression requires assembly and activation of the RNA polymerase II initiation complex at promoters of expressed genes [Zabidi et al., 2014]. Both assembly and activation are tightly regulated processes that directly control gene expression. Assembly and initiation of the RNApolII complex is influenced by the action of transcription factors (TFs) that bind the promoter DNA sequence and either recruit or activate other factors or the initiation complex.

Different transcription factors have distinct DNA sequence specificities (motifs) that are the direct result of the amino acid sequence of their DNA binding domains (DBDs) . Over 80 DBDs are known, but the specificity of the majority of those remains unknown [Weirauch et al., 2014]. Identification of the binding motif of TFs can be performed via computational analysis (motif finding) of the DNA sequences to which they bind [Bailey et al., 2009].

The actions of TFs are regulated by control of their expression as well as post-transcriptional modifications. Regulation of TFs by other TFs gives rise to transcriptional networks, whereby a master transcriptional regulator regulates other transcription factors that directly or indirectly control the expression of the effector proteins that give rise to the cellular effect. Transcription factors are also regulated by post-translational modifications, that can alter their activity either by changing their cellular localisation or by directly activating or de-activating them. Such modifications include SUMOylation, methylation, Ubiquitination, phosphorylation, acetylation and the binding of non-coding RNAs [Kim and Kim, 2014] [Bogachek et al., 2014].

Regulation of expression is not the only mode of control of gene action. Similarly to the regulation of transcription factors presented above, the action of genes is regulated by post-translational modifications of their protein products [Doll and Burlingame, 2015].

1.3 Cardiovascular Development

Cardiovascular development is paramount to the development of the embryo as failure of formation of vessels or the heart leads to early embryonic lethality. Formation of the vasculature is closely linked with blood formation and occurs in two distinct processes outlined below. Formation of the heart is also tightly linked to the development of the great vessels and the lungs in a temporal and spatial manner. Although a number of regulators of heart patterning and vessel formation are known, it is increasingly anticipated that epigenetic processes will be found to have a significant role in the development, and later homeostasis of these tissues [Chang and Bruneau, 2012].

1.3.1 Formation of the Vasculature

Vascular development in the embryo can occur by two different processes: vasculogenesis and angiogenesis. Vasculogenesis involves *de novo* formation of blood vessels by differentiation of mesodermal cells into endothelium in the yolk sac and in the developing embryo proper. In contrast, angiogenesis refers to the process of formation of new blood vessels by extension of existing ones [Baldwin, 1996]. The two processes contribute in variable degrees to the formation of vessels in different organs.

Vasculogenesis initially occurs in the yolk sac in cell aggregates termed blood islands and later in other tissues such as the liver. After formation of blood islands, the cells

that comprise each island differentiate with the outer area forming epithelial (marked by the presence of PECAM-1/CD31) cells and the interior forming blood precursors [Choi et al., 1998]. This pattern of differentiation originally suggested the existence of a common progenitor between the vascular endothelial cells and blood, and this precursor (the hemangioblast) has now been identified [Choi et al., 1998] [Nishikawa et al., 1998].

1.3.2 Development of the Heart

Heart formation is initiated in early development, during gastrulation (E6.5-E7 in the mouse) [Tam and Loebel, 2007]. Myocardial progenitor cells can be traced to a population of cells in the epiblast [Abu-Issa and Kirby, 2007]. A mesodermal population of cells that will become the adult heart after migrating through the primitive streak forms a structure known as the cardiac crescent near the head of the embryo [Moorman et al., 2003] (Figure 1.2 A1).

At this stage the endocardium arises as a plexus in the region of the cardiac crescent (Figure 1.2 B1). The exact origin of the endocardium remains unresolved (see Section 1.3.3 below). Traditional texts suggest that the cardiac crescent forms one endocardial tube on each side of the embryo, and the endocardium delaminates from them following downregulation of N-cadherin [Linask, 1992] and forms endocardial tubes in the lumen of the myocardial tubes [Abu-Issa and Kirby, 2007] [Baldwin, 1996] [Gilbert, 2003], whereas other literature suggests that the endocardial plexus is formed independently [Moorman et al., 2003]. The uncertainty in spatial origin of the endocardium reflects the uncertainty in its cellular origin.

Irrespective of the origin of the endocardium, the heart moves posteriorly as part of general rearrangements and the left and right endocardial tubes fuse in an anterior to posterior fashion, while being surrounded by the myocardium[2] which also fuses in a similar manner (Figure 1.2 C1 through C3). This leads to the formation of linear heart tube continuous with the aorta and the cardinal veins. Early events in heart formation can be visualised *in vivo* via the Nkx2-5 master heart regulator, which is expressed in the early heart and continues to be expressed in the adult [Abu-Issa and Kirby, 2007]. Later events can also be visualised by myosin heavy chain (MHC) staining [Moorman et al., 2003]. Cells that contribute to the initial stages of heart development are termed the first heart field. At a later stage, $Isl1^+$ cells from the second heart field migrate to

[2]Continuity with the mesorderm is preserved in the dorsal side in what will develop to become the mediastinum.

the cardiac region and contribute to the cranial end of the developing heart [Cai et al., 2003] [Domínguez et al., 2012]. Later during development the neural crest also contributes to heart development [Lepore et al., 2006].

The two heart fields have different contributions to the adult heart with the first heart field forming the left ventricle (LV) as well as parts of the right ventricle (RV) and atria [Chong et al., 2014]. The second heart field also contributes to the RV, the atria and the outflow track (OFT) [Abu-Issa and Kirby, 2007] [Dyer and Kirby, 2009] [Harmon and Nakano, 2013] .

At approximately embryonic day of development 11 (E11) the cardiac tube undergoes rotation (Figure 1.2 D1). Cardiac rotation is initiated when the mediastinal structure that connects the anterior section of the cardiac tube to the back of the embryo becomes disrupted and frees the cardiac tube, which rotates under its own tension [Moorman et al., 2003].

While the cardiac rotation is still occurring, the cardiac chambers begin to form by septation (Figure 1.2 E1) of the lumen of the heart, expansion of the apical portion of the heart and formation of the new myocardium [Moorman et al., 2003]. The ventricles originate from the ventricular loop [Moorman et al., 2003].

The ventricular wall forms multiple disconnected streaks at this stage giving it a 'spongy' appearance, in a process termed trabeculation. Compaction later in development will fuse these trabeculae into a more uniform and compact muscle layer.

Valve Formation

Valve formation commences at E9.5 with the excretion of extracellular matrix (ECM) by the myocardium of the heart in the location of the future valves [DeLaughter et al., 2011] in the OFT and atrioventricular canal (AVC) regions, leading to the formation of the cardiac cushions [Chakraborty et al., 2010] [Hinton and Yutzey, 2011]. The cardiac cushions will act as primitive valves in the developing heart, establishing a unidirectional blood flow, albeit with some regurgitation [Wu et al., 2013].

Following cardiac cushion formation, a subpopulation of the endocardial cells (the inner layer of the heart, Section 1.3.3) overlaying the cushions undergo endothelial-to-mesenchymal transformation (EMT). EMT occurs under the control of *Erg* [Vijayaraj et al., 2012] [Wu et al., 2013], a member of the ETS family of proteins. During EMT, endocardial cells delaminate from the endothelium and invade the underlying cardiac

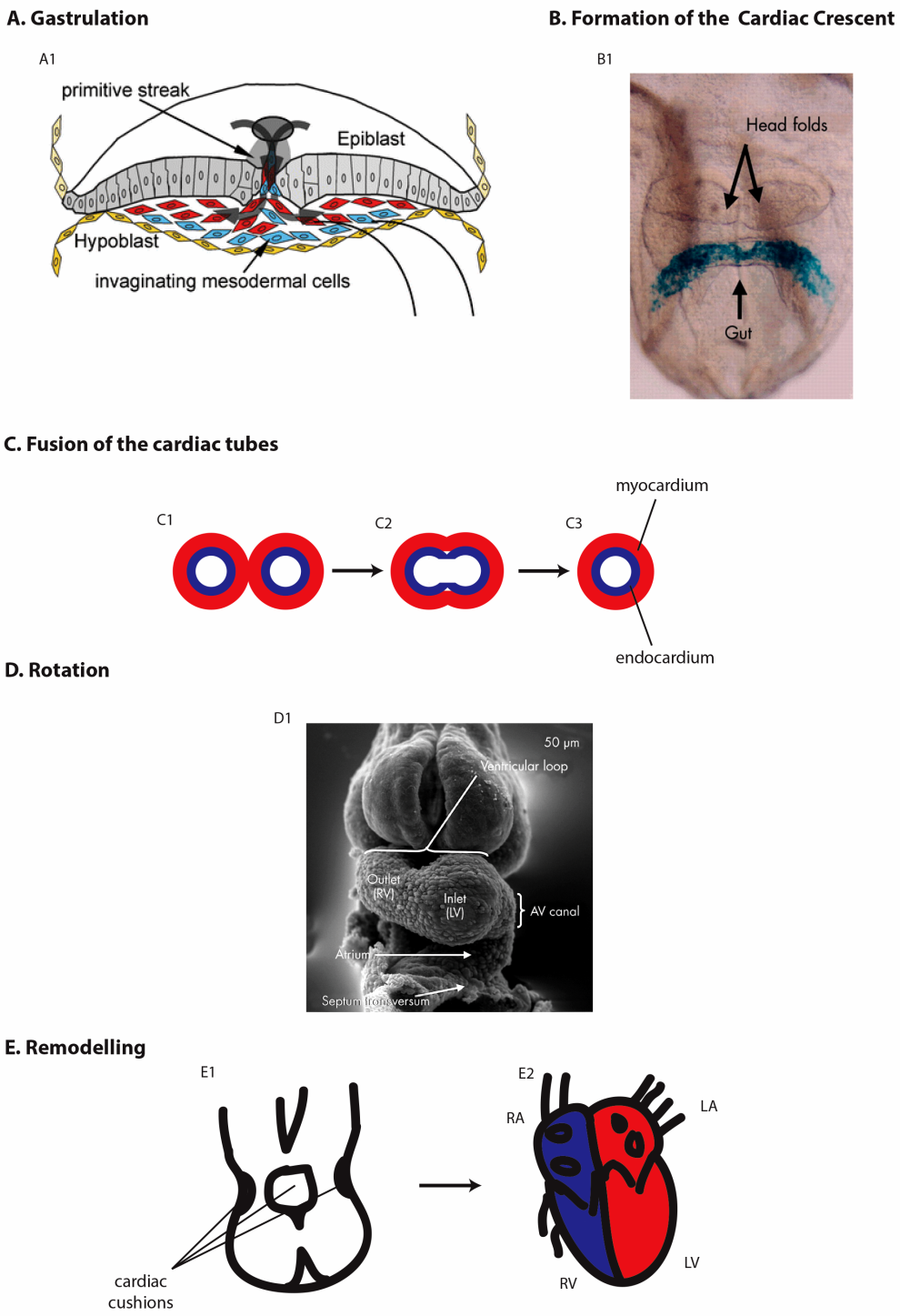


Figure 1.2: Diagramatic summary of the embryonic development of the heart. The heart is formed from anterior lateral plate mesoderm cells that migrate through the primitive streak during gastrulation (A1), adapted from [Harris and Black, 2010]. Cardiac progenitors form the cardiac crescent, here shown labelled with alpha-myocin (B1), adapted from [Moorman et al., 2003]. The cardiac crescent forms two concentric tubes on each side of the embryo (C1). The outer myocardial tubes fuse first followed by the inner endocardial tubes (C2) into a single heart tube in the midline. (C3). The cardiac tube comprises of an outer myocardial layer and and inner endocardial (C3). The heart tube undergoes rotation (D1), adapted from [Moorman et al., 2003]. The heart remodels through the formation of cardiac cushions (E1) and rearrangement of inflow and outflow tracks into the adult four chamber structure (E2) with separate right atrium (RA), right ventricle (RV), left atrium (LV), left ventricle (LV).

jelly where they remodel the ECM. Initiation of EMT and valve formation requires both myocardium and endocardium from the valvular regions [Baldwin, 1996] [Chakraborty et al., 2010]. Transplantation experiments have shown that replacement of either the endocardium or myocardium by respective tissue from other parts of the heart ablates valve formation [Mjaatvedt et al., 1987], demonstrating the spatial heterogeneity of the endocardium. A portion of the population of endocardial cells overlaying the future valve region do not however undergo EMT. These cells remain in the epithelial layer and proliferate to generate the valve leaflets [Wu et al., 2013].

The process of EMT is well studied due to the existence of a simple assay that allows it to be recapitulated *in vitro*. Unlike EMT however, the mechanism of cardiac cushion remodelling is poorly understood because no *in vitro* system exists to recapitulate these events [de Vlaming et al., 2012].

1.3.3 Origin and Differentiation of the Endocardium

The endocardium is a layer of endothelial cells lining the internal surfaces of the ventricles and atria of the heart. The endocardium first appears shortly after gastrulation almost concurrently with the formation of the cardiac crescent. Future endocardial cells down-regulate expression of the adhesion molecule N-cadherin and separate from the rest of the cardiogenic mesoderm. The close spatial relationship of the endocardial and myocardial populations suggests a common developmental precursor [Baldwin, 1996] as opposed to a vascular endothelial precursor. The exact origin of the endocardium however remains the subject of debate. Conflicting reports exist in the literature with some studies suggesting a vascular origin [Milgrom-Hoffman et al., 2011] [Harris and Black, 2010] and others a common origin with other cardiac cells [Misfeldt et al., 2009] [Moretti et al., 2006] [Kattman et al., 2006]. Furthermore, it has been suggested that endocardial cells may be able to form vasculature [Wu et al., 2012], further complicating our understanding of the endocardium.

A multipotent progenitor originating from the second heart field ($Isl1^+$) that can differentiate into myocardium, smooth muscle and endothelial cells was described in ES cell culture in 2006 by Moretti and colleagues [Moretti et al., 2006]. This cell line was termed a multipotent, cardiovascular progenitor (MICP) and was defined by the $Isl1^+$/$Nkx2\text{-}5^+$/$Flk1^+$ expression signature and its existence in the developing embryo was confirmed.

Around the same time Kattman and colleagues, demonstrated the existence of an

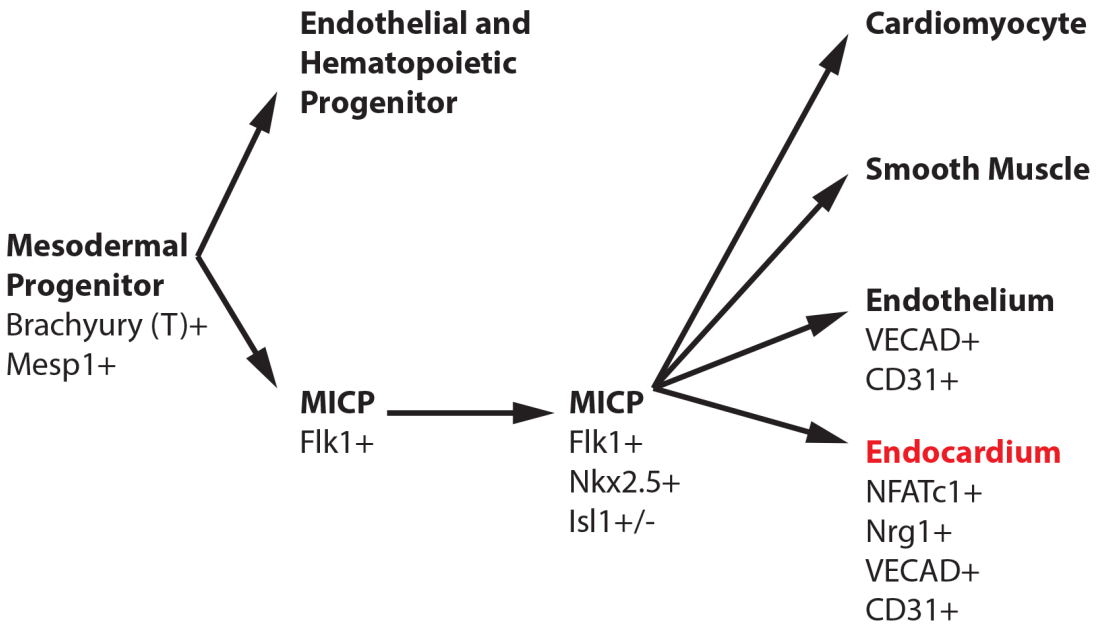


Figure 1.3: Working model of the embryonic origin of the endocardium. The endocardium originates from a multipotent cardiovascular progenitor population that can also give rise to other cardiac tissues. (Adapted from [DeLaughter et al., 2011]).

even more general Flk1$^+$ cardiovascular progenitor with endothelial, cardiomyocyte and vascular smooth muscle lineages [Kattman et al., 2006]. The relationship between the two populations is not entirely clear but a working model [DeLaughter et al., 2011] presents this population as a predecessor of the MICP (See Figure 1.3).

A unique endocardial identity and origin is however supported by the existence of a distinct endocardial marker [de la Pompa et al., 1998] [Ranger et al., 1998] (see following Section) as well as the existence of a double gene knock-out that shows specific endocardial defects. The single knock-out of either of the two structurally related tyrosine kinase receptors Tie1 or Tie2 (Tek) expressed throughout the endothelium is embryonically lethal, but does not show any defects in the initial formation of the vasculature or of the endocardium. In contrast, the double mutant shows specific failure of the initial formation of the endocardium suggesting a distinct identity for this cardiac population [Puri et al., 1999].

1.3.4 Congenital Heart Defects

Congenital heart defects (CHD) are the most common birth defects affecting between 1% to 5% of live births [Leirgul et al., 2014] [Hoffman and Kaplan, 2002]. CHD is subclassified into a number of individual defects with different management. These include patent ductus arterious (failure of the ductus arteriole to close after birth), ventricular

septal defects (defects in the septum separating the two ventricles), atrial septal defects, pulmonic and aortic stenosis, coarctation of the aorta (narrowing of the aorta at the ductus arteriole, tetralogy of fallot (a complex malformation characterised by pulmonary infundibular stenosis, overriding aorta, ventricular septal defects and right ventricular hypertrophy), complete transposition of the great arteries (where the right ventricle is connected to the aorta and the left ventricle to the pulmonary artery), hypoplastic left and right heart syndromes, persistent truncus arteriosus (a ventricular septal defect), single ventricle and total anomalous pulmonary venous connection (whereby the pulmonary veins connect to the systemic venous circulation) [Hoffman et al., 2004].

A subclass of congenital heart disease is congenital valvular disease, whereby the cardiac valves fail to develop correctly. Congenital valve disease encompases bicuspid aortic valve defects, mitral valve prolapse, pulmonary valve stenosis and Ebstein anomaly of the tricuspid valve [LaHaye et al., 2014] [Lincoln and Garg, 2014].

Despite the high incidence of valve defects, treatments for this class of CHDs are not effective due to lack of understanding of the mechanisms that lead to these malformations [Lincoln and Garg, 2014]. Treatment is limited to surgical replacement with mechanical or bioprosthetic valves. However such treatments have severe caveats such as the limited durability and need for continuous coagulation [LaHaye et al., 2014].

Genetic basis for some of these malformations has been described suggesting a genetic basis for other valvular malformations. Bicuspid aortic valve defects (one of the most common valve malformations present in 1-2% of the population) have been linked to NOTCH1 [McBride et al., 2008] in human and *Gata5* mutations in the mouse [Bonachea et al., 2014]. Mitral valve prolapse has been functionally associated with Marfan syndrome which has well established genetic aetiology [LaHaye et al., 2014].

Improvements in clinical management have improved the prognosis of such defects resulting in a large number of adults living with valvular CHD [Hoffman et al., 2004] posing significant challenges in the management of resulting complications [Tutarel, 2014] and further necessitating the understanding of the molecular aetiology of these malformations to improve treatment. Furthermore, detailed understanding of the development of valves may allow *in vitro* growing of valves from patient-derived embryonic stem cells in the future.

Significance of the Endocardium and its Role in Congenital Heart Defects

Beyond its primary role in the formation of an endothelial layer in the heart lumen, the endocardium has other functions [Harris and Black, 2010]. Specifically, endocardium is required for cardiac development and in particular trabeculation of the myocardium, differentiation of myocytes into Purkinje conduction fibers, formation of cardiac valves and separation of the OFT into the pulmonary artery and aorta.

The endocardium is involved in the formation of cardiac valves by partially undergoing EMT at the sites of the future valves [Wu et al., 2013] [von Gise and Pu, 2012]. Under the control of NFATc1, a portion of endocardial cells undergo EMT and invade the cardiac cushions to form the valve mesenchyme and remodel the cardiac cushions. These endocardial cells are responsible for valve formation, whereas endocardial cells that do not undergo EMT will form the valve leaflet [Wu et al., 2013].

The endocardium has also been implicated in conduction fiber formation. The presence of endocardial cells is required for expression of Purkinje fiber marker genes, by these cells. Furthermore, ablation of *Neuregulin* a soluble signalling protein excreted by the endocardium results in conduction defects, in addition to other myocardial defects [Mikawa and Hurtado, 2007].

Ablation of *Neuregulin* also results in compaction and trabeculation defects exemplifying the role of the endocardium in these processes. Finally, the endocardial specific knockout of *Fkbp1a*, a cis-trans peptidyl-prolyl isomerase results in non-compaction of the myocardium [Chen et al., 2013].

In addition to the above, the endocardium has recently been identified as a source of a population of cells leading to the formation of the coronary circulation during trabeculation [Tian et al., 2014] directly implicating these cells in cardiac muscle vascularisation.

Given the diverse roles of the endocardium in cardiac development and the clinical significance of CHDs, understanding the regulation, role and actions of the endocardium during embryonic development is critical for both understanding the aetiology and improving the management of these defects.

NFATc1 as a Marker and Functional Component of Endocardial Cells

Nuclear Factor of Activated T-cell (NFATc1), a calcineurin dependent transcription factor, was identified as an early marker of endocardial cell differentiation in two simultaneous publications in 1998 [Ranger et al., 1998] [de la Pompa et al., 1998].

Between E8.5 and E10.5 NFATc1 expression is confined to the developing endocardium [Misfeldt et al., 2009]. The expression is transient and later during development it is expressed in other embryonic tissues such as limb cartilage and hair follicles [Misfeldt et al., 2009], while in the adult NFATc1 is involved in activation of lymphocytes [de la Pompa et al., 1998]. At E9.5 and until E11.5 NFATc1 expression is upregulated in the OFT and AVC endocardial cells and downregulated in the rest of the endocardium [Wu et al., 2013] [Misfeldt et al., 2009].

At the time of identification of NFATc1 as an endocardial marker it was recognised that it was also essential for valve formation as its ablation leads to valvular defects [Ranger et al., 1998] [de la Pompa et al., 1998]. It was later appreciated however that although essential for valvulogenesis NFATc1 is not essential for commencement of EMT at the site of the future valves, thus making its exact role unclear.

More recent work supports the notion that the role of NFATc1 is to block rather than induce EMT in endocardial cells [Wu et al., 2013]. This blockage is proposed to occur in only a portion of endocardial cells thus being the deciding factor to allocate cells between the valve leaflet formation and valve mesenchyme. Specifically, evidence exists to support the idea that downregulation of NFATc1 is responsible for the commencement of EMT and in fact its presence arrests cells in a pre-EMT state at E10.5 by suppression of Snail1 and Snail2 regulators [Wu et al., 2011a]. The way in which NFATc1 levels remain high in only a fraction of endocardial cells remains unknown.

In addition to the aforementioned role, expression of NFATc1 is proposed to promote valve elongation [Wu et al., 2013]. It is known that NFATc1 can autoamplify itself and thus maintain persistent expression once activated and it appears that this auto-activation is important in the promotion of valve elongation [Zhou et al., 2005]. The mechanism by which NFATc1 is eventually downregulated remains unclear.

Regardless of its role in the regulation of EMT and its importance in valve development, NFATc1 serves as an excellent marker of the endocardium at E9.5 (Figure 1.4). A NFATc1 reporter transgene that recapitulates the endogenous expression pattern has been made by insertion of a lacZ reporter 400 bp upstream of the first exon of NFATc1 on a extra-chromosomal bacterial artificial chromosome containing the first two exons of NFATc1 and upstream enhancer sequence [Misfeldt et al., 2009]. The BAC locus does not entail the entirety of the NFATc1 transcript and cannot result in increased transcription of the full-length transcript. Replacement of the lacZ reporter in this construct by a dual

reporter [de Felipe and Ryan, 2004] including mCherry and Cerulean allows separation of the endocardial cells via flow cytometry.

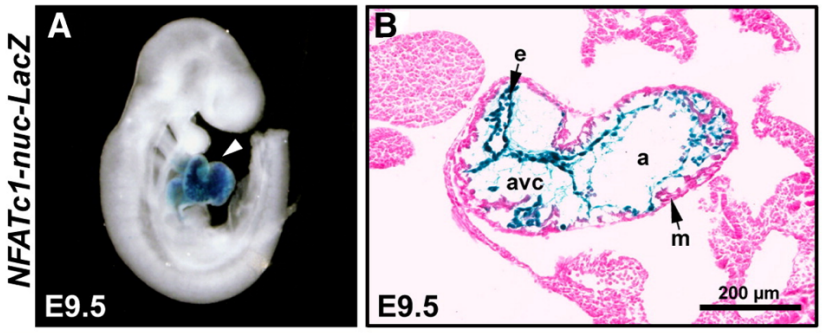


Figure 1.4: NFATc1, here transgenically labelled with lacZ, serves as an excellent marker of the endocardium at E9.5. Reproduced from [Misfeldt et al., 2009].

1.3.5 Recapitulation of Endocardial Development by Embryoid Bodies

Differentiation of embryonic stem cells in suspension culture leads to the development of structures called embryoid bodies [Smith, 2001]. Embryoid bodies have been found to be able to differentiate into multiple different cell types and have been successfully used for the identification of the embryonic precursor of the hematopoietic and endothelial cell lines, the hemangioblast [Choi et al., 1998].

In 2007, Narumiya and colleagues demonstrated that the protocols that direct differentiation of embryoid bodies towards cardiac cell lines also lead to the development of endocardial-like cells [Narumiya et al., 2007]. At the time these endocardial cells in embryoid bodies were found to have a consistent expression profile with *in vivo* endocardium and mesodermal differentiation. Furthermore, they were found to co-localise with myocardial cells and no endocardial differentiation was observed in the absence of myocardium. This observation is consistent with the reciprocal interactions between the endocardium and myocardium that are required for cardiac development.

Prior to this work, two reports outlining the existence of cardiovascular precursors had been published [Kattman et al., 2006] [Moretti et al., 2006]. These studies showed that the cardiovascular progenitors could give rise to endothelial cells among other cardiac populations, but their specific endocardial identity had not been established because of the absence of a reliable marker.

In 2009, Misfeldt and colleagues, demonstrated the presence of NFATc1$^+$/CD31$^+$ cells in embryoid bodies with the use of a labelled NFATc1 construct and immunohistochemical

analysis [Misfeldt et al., 2009]. Furthermore, in the same study a temporal expression pattern of NFATc1 and other mesodermal as well as cardiac differentiation markers consistent with mesodermal endocardial differentiation and close spatial association of endocardial and endothelial cells in embryoid bodies were observed.

1.3.6 Epigenetics in Heart Development and Remodelling

Over the past few years it has become clear that epigenetic processes play a significant role in the strict spatio-temporal coordination required for the development, maintenance and remodelling of the heart. Although many genes are known to be involved in the epigenetic regulation of heart development and remodelling, the exact mechanistic details of their actions remain largely unknown. This in conjunction with recent developments that have made whole-genome epigenetic interrogation possible (Section 2.2) have sparked great interest in the epigenetic processes that regulate the heart and cardiovascular disease [Baccarelli et al., 2010].

An extensive body of evidence supporting the role of epigenetic processes in cardiac development, physiological remodelling and disease exists and examples are presented here.

The basic question of epigenetic heterogeneity of heart tissue has been demonstrated by identifying epigenetic differences between the RV and LV via small scale qPCR [Mathiyalagan et al., 2010]. Larger studies have examined the evolution of chromatin marks during cardiac cell specification and have identified novel regulators of cardiac cell specification (such as *Meis2*) [Paige et al., 2012] [Wamstad et al., 2012] . These studies have demonstrated whole genome epigenetic analysis is a viable methodology for identification of cardiac regulators and have shown that temporal epigenetic changes can be used to discriminate regulatory from constitutively expressed cardiac genes [Paige et al., 2012].

The utility of whole-genome approaches in elucidating heart specific epigenetic approaches that have been refractory to more classical analysis was shown by Blow and colleagues by performing p300 histone acetyl-transferase (HAT) specific ChIP-seq in hearts from E11.5 mouse embryos [Blow et al., 2010]. In this study the investigators were able to identify over 3,000 candidate heart enhancers that are otherwise poorly conserved and demonstrate that most are functional by generating and characterising transgenic mice for a subset of them.

Furthermore, specific epigenetic regulators have been linked with cardiac develop-

ment. *Brg1*, a subunit of the BAF complex, controls temporal and spatial expression of *Adamts1* metaloprotease in the endocardium and myocardium to control extracellular matrix deposition. It also directly controls gene expression in the myocardium, with its deletion leading to temporally specific cardiac defects [Chang and Bruneau, 2012]. *Brg1* can control the switch of MHC isoform expression in adult hypertrophic cardiomyopathy from the adult isoform to the fetal isoform [Chang and Bruneau, 2012]. It is well established that during cardiac hypertrophy the adult αMHC is down-regulated and the fetal isoform βMHC is upregulated. Multiple Histone Deacetylase (HDAC) mutations have also been associated with different forms of cardiac hypertrophy [Tingare et al., 2013]. Cardiac failure and hypertrophy have also been more directly associated with H3K4me3 changes [Kaneda et al., 2009].

A number of associations between epigenetic factors and human heart defects have been found. For example, Wolf-Hirschhorn syndrome is characterised by growth retardation, craniofacial malformations, learning disabilities and heart defects and occurs as a result of a deletion of chromosome 4q16.3, which includes a gene coding for a H3K36 methyltransferase. H3K36 methyltransferase ablation results in atrial and ventricular septal defect [Vallaster et al., 2012]. The CHARGE syndrome, characterised by heart defects among other abnormalities is closely associated with genetic mutations of the *Chd7* gene, a member of the chromodomain ATP-dependent chromatin modifiers [Vallaster et al., 2012].

Collectively, the above indicate that epigenetic processes may have a significant and clinically relevant role in heart development and provide justification for further studies into the epigenetic processes that act during heart development.

1.3.7 Epigenetics of the Vascular Endothelium

Epigenetic processes are also known to regulate gene expression in the vascular endothelium in a number of loci [Yan et al., 2010].

Nos3 is the endothelial subtype of Nitric Oxide (NO) synthetase. NO is a molecule critical for endothelial regulation and signalling. *Nos3* expression has been found to be epigenetically regulated. Initial evidence for epigenetic regulation of *Nos3* originated from a transgene expressed in all cell types irrespectively constitutive expression of *Nos3* in that cell type. It was however observed that the expression of the transgene could be controlled by methylation [Krause et al., 2013], suggesting an epigenetic mechanism for

constitutive regulation of the gene. The importance DNA methylation [Chan et al., 2004] and histone modifications [Fish et al., 2005] in *Nos3* regulation has been confirmed in *in vivo*. In agreement with the above *Nos3* has been found to be epigenetically regulated in human umbilical endothelium [Krause et al., 2013].

Epigenetic gene regulation in the endothelium has also been demonstrated in the *Notch4* locus. The Notch signalling pathway is a critical pathway in angiogenic vascular remodelling and *Notch4* is preferentially expressed in endothelial cell. Cell-type specific histone modifications have been found to play a role in regulation of *Notch4* [Wu et al., 2005].

The promoter of VWF has also been found to be regulated by recruitment of HDACs by NFY [Peng and Jahroudi, 2003] and E-selectin expression in response to TNF induction has been associated with histone hyperacetylation, phosphorylation, and methylation [Edelstein et al., 2005]. Finally, *Robo4*, an endothelial specific protein, is known to be regulated by differential methylation of its promoter by specific demethylation during development [Okada et al., 2014].

Overall the above locus-specific evidence point towards a greater genome-wide role for epigenetic regulation of gene expression in the endothelium.

1.3.8 Epigenetics of the Endocardium

The contribution of epigenetic processes in the developing endocardium has not been investigated. However, given that the endocardial tissue is an endothelial tissue with possible cardiac origin and that epigenetic processes are known to regulate gene expression in both endothelium and heart, it is highly likely that the same processes also have a role in endocardial development and differentiation.

To date transcriptional regulators that are unique to the endocardium have not been identified and many of the transcriptional regulators identified in this tissue are shared with the endothelium. Furthermore, the endocardium is largely morphologically identical to the endothelium (Prof. Scott Baldwin, personal communication) during development and exhibit morphological differentiation later in development, suggesting the presence of latent information that manifests later in development, compatible with a distinct epigenetic identity for this population.

Given the above, further investigation into the epigenetics of the endocardium is warranted.

1.4 Imprinting

Imprinting refers to the parent-of-origin specific expression of some genes in mammals and flowering plants [Reik and Walter, 2001] (Figure 1.5). Imprinting was first described in the early 1980s after pronuclear transplantation technology allowed the generation of embryos with uniparental disomies (UPDs). The examination of these embryos revealed that the effect of some UPDs depended on their origin strongly suggesting the non-equivalence of the parental genomes [Surani et al., 1984] [McGrath and Solter, 1984]. Since these initial discoveries parent-of-origin specific expression of approximately 150 in the mouse and 72 genes in human has been described [Williamson et al., 2014] [Morison et al., 2005].

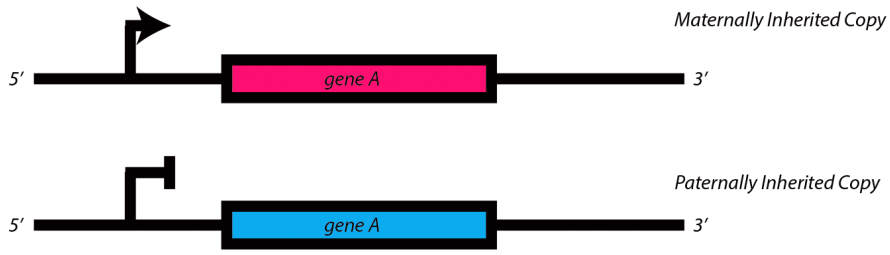


Figure 1.5: Example of a hypothetical maternally imprinted locus, only the maternally inherited copy is expressed; the paternally inherited copy is silenced.

1.4.1 Functional Significance of Imprinting

The functional significance of imprinting remains unresolved [Patten et al., 2014]. Several theories have been proposed to explain this phenomenon but no theory has been conclusively shown to account for all the known imprinted loci. It is possible that different explanations can account for imprinting at different loci.

Theory of Coadaptation

The theory of coadaptation suggests that imprinting has evolved as a result of genetic compatibility or incompatibility of loci between the mother and the offspring leading to selective abortion and/or implantation of embryos [Wolf and Hager, 2006] [Wolf and Hager, 2009] [Wolf and Brandvain, 2014].

The theory proposes that when matching alleles between the mother and the offspring are beneficial, the paternal allele (which may be incompatible with the maternal allele) is silenced. This situation may for example arise in loci the protein product of which may result in an immune response from the mother against the foetus. Conversely, according

to the theory of coadaptation, when it is beneficial for the maternal and paternal alleles to mismatch the maternal allele (that will match at least one of the maternal allele) is silenced. In either of this type of locus, imprinting can potentially increase fitness of the offspring by effectively 'hiding' from the mother the allele that has the potential to reduce fitness.

The coadaptation theory can account for the excess of maternally imprinted genes in comparison to the paternally imprinted genes. Furthermore, the theory can potentially account for an excess of maternally expressed genes in the placenta in mammals and in early seed development in endosperms.

Theory of Parental Conflict

The theory of parental conflict suggests that imprinting has evolved as a result of conflict between the paternal and maternal genomes [Moore and Haig, 1991]. According to this theory, the paternal genes have an evolutionary advantage when they promote the transfer of nutrients from the mother to the offspring and the maternal genes have an advantage when they limit this transfer. The theory of parental conflict is based on the observations that androgenetic mouse embryos show limited embryo development but excess extra-embryonic tissues, whereas the reverse pattern is observed in parthenogenetic embryos. Furthermore, a similar pattern is observed in humans with some UPDs, the most well described being the deletion of 15q11-13, the maternal loss of which gives rise to Angelman syndrome and the paternal loss giving rise to Prader-Willi syndrome (see Section 1.4.2 below)

The theory of parental conflict is further supported by the imprinting pattern of *Igf2* and its receptor *Igf2r* in the mouse. *Igf2* encodes for an insulin-like growth factor that is embryonically expressed and is involved in nutrient transfer between the mother and the foetus. The *Igf2* gene and its receptor exhibit opposing imprinting patterns with *Igf2* maternally expressed and *Igf2r* paternally expressed. *Igf2*, that is widely paternally expressed including in extraembryonic tissues, promotes embryo growth (and transfer of nutrients from the mother to the embryo); accordingly paternal *Igf2* mutants display smaller size at gestation. In contrast, the Igf2r receptor, one of the receptors for Igf2, which may sequester and ameliorate the action of Igf2 in other receptors, is maternally expressed [Halg and Graham, 1991].

1.4.2 Imprinting in the Brain

Many imprinted genes are highly expressed in the brain or spinal cord [Davies et al., 2005] and the importance of imprinting mechanisms in the brain and its development has been highlighted by a number of separate lines of evidence.

Imprinting has been implicated in a number of syndromes that affect brain development including Angelman and Prader-Willi syndromes [Keverne, 1997]. Prader-Willi syndrome is characterised by mild mental retardation, obesity, hypotonia, hypothalamic disfunction and psychoses [Davies et al., 2005]. Paternal selection of the 15q11-13 locus results in Prader-Willi syndrome, but in contrast, maternal deletion of the locus results in Angelman syndrome [Lewis et al., 2014]. Angelman syndrome is also characterised by mental retardation, but also movement disorders, easily provoked laughter and speech difficulties [Larson et al., 2014].

Furthermore, specific imprinted genes are implicated in brain development and function. The imprinted *Peg3* gene, has been implicated in maternal care behaviour [Murphy et al., 2001] [Keverne et al., 1996] and polymorphisms in this gene have been associated with different maternal care in mice strains [Chiavegatto et al., 2012]. Furthermore, duplication of the maternal but not paternal *Nnat* gene results in aberrant cerebral development.

Further evidence for the importance of imprinting in the developing brain originates from the distinct contributions of androgenetic and parthenogenetic cells in brain chimeric embryos. Androgenetic cells, containing only a chromosomal compliment from the father, contribute primarily to the hypothalamus, whereas parthenogenetic cells contribute to the cortex and hippocampus but not the hypothalamus. In addition, parthenogenetic cells enhance brain growth whereas androgenetic cells diminish it [Keverne et al., 1996].

In the mouse, at post-natal day 21 (P21), the post-natal growth spurt has been largely completed and the reduction in neuronal numbers that is observed in the adult at (P50) has not yet occurred. At this time point, the neurons make up approximately 50% of the total brain cells and glia account for approximately 48%, with other cell types such as oligodendrocytes accounting for part of the remainder [Lyck et al., 2007].

Collectively, the above support an extensive role for imprinting in the brain and make this tissue suitable for investigation of this phenomenon.

1.4.3 Imprinting and Epigenetic Gene Regulation

At the molecular level imprinting has been shown to be mechanistically underpinned by the differential methylation of the two parental alleles that is established during gametogenesis. Furthermore, methylation marks responsible for imprinting escape large scale resetting of the methylome during implantation and later development [Hajkova et al., 2008] [Hajkova, 2010] [Cowley and Oakey, 2012], leading to differentially methylated regions in the adult organism.

An important distinction between imprinting control regions (ICRs) and differentially methylated regions (DMRs) must be made at this point. DMRs comprise all the genomic locations that are differentially methylated between the two parental alleles and include regions that have no effect on gene expression and regions of which the methylation pattern is not established in gametes. In contrast, ICRs are regions which are defined by knockout technology in mice and which when ablated have been shown to control the imprinting of defined genes.

Unlike methylation at most loci, where methylation is positively correlated with the repression of gene expression, DNA methylation at imprinted loci may positively correlate with activation or repression of genes as the mechanisms of translation of the methylation patterns into parent-of-origin specific expression are diverse. These mechanisms include promoter methylation, antisense transcript expression and the boundary element establishment, as is the case for the *H19/Igf2* locus (Section 1.5.3).

1.5 CTCF

The CCCTC-binding factor (CTCF), a 727 amino acid protein, was first identified as one of two regulators that bound a poorly conserved 50-60 bp sequence upstream of the chicken c-myc gene, that contained the CCCTC central motif sequence [Lobanenkov et al., 1990]. It was cloned and further characterised as a highly conserved multivalent transcriptional repressor six years later [Filippova et al., 1996]. At the time it was appreciated that it can bind heterologous DNA sequences via the combinatorial action of its Zinc-finger domains. Concurrently, it was independently discovered as a silencer of the chicken lysozyme gene and named NeP1 [Baniahmad et al., 1990]. It was not until the gene was cloned and sequenced that it was realised that NeP1 and CTCF were the same protein [Burcin et al., 1997].

CTCF is ubiquitously expressed and highly conserved protein with near 100% amino acid sequence identify between human, mouse and chicken and furthermore, its depletion is embryonically lethal prior to implantation, suggesting an important role in cell maintenance [Phillips and Corces, 2009] .

CTCF has been shown to have insulator function by binding to boundary elements between enhancers and promoters and abrogating expression in enhancer blocking assays [Bell et al., 1999]. As a result of this work, CTCF is now widely considered to be primarily an insulator [Phillips and Corces, 2009]. Other evidence however suggests CTCF involvement in a diverse repertoire of functions in addition to insulation. Specifically, CTCF has been implicated in transcriptional regulation [Vostrov and Quitschke, 1997] [Kuzmin et al., 2005], X chromosome inactivation, large scale organisation of the genome via looping, V(D)J recombination in lymphocytes [Guo et al., 2011a], association with lamina-associated domains and imprinting [Phillips and Corces, 2009].

More recently CTCF has been implicated in the evolution of genomic organisation via propagation of its novel binding sites via retrotransposition [Schmidt et al., 2012]. The same study revealed a previously unappreciated second 9 bp binding motif at a consistent spacing from the canonical CCCTC binding motif that correlates with stronger binding and better conservation across species.

1.5.1 Insulator Role of CTCF

As aforementioned, CTCF is best known as an insulator protein that prevents the action of activating sequence elements extending past its binding site. The role of CTCF in enhancer insulation was first proposed in 1999 based on work on the β-globin locus demonstrating the existence of a 42 bp sequence element that is necessary for enhancer blocking and that concurrently binds CTCF [Bell et al., 1999].

The insulator role of CTCF was later described in more detail at the *H19/Igf2* locus, where it is responsible for imprinting via insulation of enhancer action in a parent-of-origin allele-specific manner. CTCF has been shown to demarcate regulator domains and abrogate correlation of expression of nearby genes in a genome-wide manner [Xie et al., 2007], consistent with a genome-wide insulator role.

The mechanism of action of CTCF as an insulator is unknown, although it is hypothesised that insulation involves formation of loops via dimerisation [Phillips and Corces, 2009]. This raises the possibility that CTCF may only be necessary for the genomic or-

ganisation required or associated with enhancer blocking activity and does not directly act as an insulator.

1.5.2 Transcriptional Activation Role of CTCF

CTCF binding has been shown to be activating in at least two loci, *Irak2* [Kuzmin et al., 2005] and amyloid β-protein precursor promoter [Vostrov and Quitschke, 1997].

In the amyloid β-protein precursor promoter the activating function has been mapped further through the use of deletions to the C-terminus. The C-terminus is independent of the Zinc-finger DNA binding domain, suggesting a specific role in transcriptional activation. In the *Irak2* locus, CTCF has been, by deletion of its binding site, shown to play a major role in the promoter activity of the *Irak2* gene.

Despite these known instances of activation, the genomic distribution of CTCF is dissimilar to that of canonical transcription factors and only a small portion of CTCF binding sites are in the proximal (2.5 kb) promoter region of genes [Kim et al., 2007].

The mode of action of CTCF in loci where it has an activating role has not been elucidated in great detail, although limited evidence exists to suggest that it can recruit RNApolII to gene promoters [Chernukhin et al., 2007].

1.5.3 Role of CTCF in Imprinting

CTCF is part of a mechanism that connects epigenetic marks and expression in the context of imprinting. CTCF binding is known to be inhibited by prior DNA methylation and also inhibit methylation of DNA to which it is bound [Phillips and Corces, 2009], allowing it to provide a functional link between DNA methylation and control of expression. The role of CTCF in imprinting is best studied in the *H19/Igf2* locus. This locus contains a single ICR, which is differentially methylated between the two paternal alleles during gametogenesis and survives nuclear reprogramming.

On the maternal chromosome the ICR is unmethylated, whereas in the paternal it is methylated. Downstream of the ICR lies the *H19* gene locus and further downstream tissue-specific enhancers. Approximately 90 kb upstream of the ICR lies the coding region for *Igf2* (Figure 1.6) [Kanduri et al., 2000].

CTCF binds to the *H19/Igf2* ICR in a methylation sensitive way, directing the action of the downstream enhancer elements. On the paternal unmethylated copy of the locus, CTCF does not bind and the enhancers are able to activate the more distant *Igf2*

promoter. In the maternal copy, however, CTCF does bind and it limits the action of the enhancer elements to the more proximal *H19* locus. This results in maternal expression of *H19* and paternal expression of *Igf2* [Phillips and Corces, 2009]. The action of CTCF in this locus is mediated by the formation of loops on the maternal *Igf2* making it inaccessible to enhancer elements [Yoon et al., 2007].

CTCF has also been implicated in the imprinting of other loci, such as that of *Meg1/Grb10*. In this locus, CTCF binding sites have been found in the mouse promoter of the *Meg1/Grb10* transcript but not at the human homologue correlating with the different imprinted status of this gene in the two species [Hikichi et al., 2003].

Given the above roles of CTCF in the control of gene expression in response to DNA methylation and its potential role in nuclear organisation CTCF has been proposed to be part of a more general heritable epigenetic system [Phillips and Corces, 2009].

1.5.4 Role of CTCF in Genomic Organisation

A role for CTCF in genome organisation has been proposed on the basis of both its ubiquity and association with lamina-associated domains (LADs). LADs are large (0.1-1 Mb) genomic locations of low gene density and repressed expression that physically localise to the nuclear periphery. CTCF has been found in a low (10-15%) but significant proportion of the borders of LADs suggesting a role in definition of their boundaries [Guelen et al., 2008]. In addition, CTCF binding sites have been found to correlate with domains of distinct histone modifications [Handoko et al., 2011], further supporting a role in genomic organisation. Finally, CTCF mediated inter-chromosomal interactions have been found to have a role in X-chromosome inactivation [Phillips and Corces, 2009].

1.6 Cohesin

Cohesin is a multimeric protein complex with a well-described role in sister chromatin cohesion from the S-phase until chromatin segregation [Onn et al., 2008]. It comprises two SMC (structural maintenance of chromosomes) subunits, Smc1 and Smc3, as well as Scc3 and Rad21 (also known as Med1). Smc1 and Smc3 can dimerise and form a 45-nm ring structure, to which Scc3 and Rad21 bind [Onn et al., 2008].

Cohesin connects chromosomes throughout their length during replication, but its action is antagonised by the Wapl protein, resulting in separation of the two sister chro-

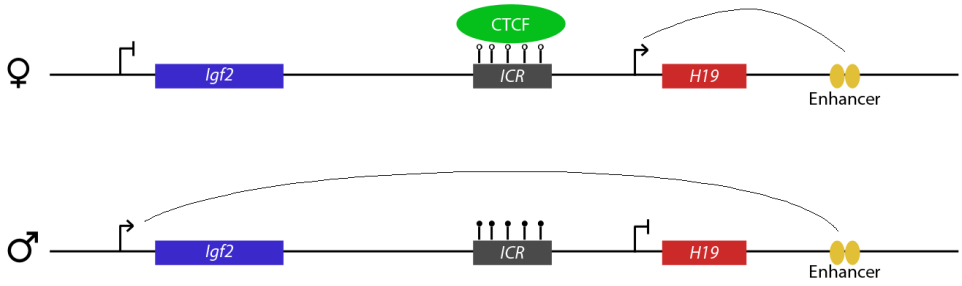


Figure 1.6: Simplified structure of the *H19/Igf2* imprinted locus. Allele-specific binding of CTCF on the differentially methylated ICR blocks the action of the enhancer element resulting in coupled imprinting of the *H19* or *Igf2* transcripts.

matids, with the exception of the centromeric region which is protected from the action of Wapl. Removal of the cohesin at the centromere, by cleavage of the Rad21 subunit, by Separase results in disassociation of the two sister chromatids during anaphase [Ocampo-Hafalla and Uhlmann, 2011].

In addition to the above well-established role, cohesin has been implicated in transcriptional regulation, DNA repair [Onn et al., 2008] and chromatin organisation [Sofueva et al., 2013]. These additional roles of cohesin are supported by the continued expression of cohesin after mitosis, including in non-proliferating cells (such as neurons) [Schmidt et al., 2010] and the existence of mutants that show minimal replication defects but give rise to a class of severe developmental conditions termed cohesinopathies, that include Roberts and Cornelia de Lange Syndromes [Skibbens et al., 2013].

Cohesin has been implicated in transcriptional regulation in a number of loci, including regulation of the homeobox genes in *Drosophila melanogaster* and the expression of *Runx1* and *Runx3* genes in *Danio rerio* during development. Furthermore, cohesin has been found to associate with pluripotency transcription factors in ES cells and be essential for maintenance of pluripotency, suggesting a regulatory role. In addition, cohesin functions as a boundary element in the regulation of the silent mating cassette in yeast [Onn et al., 2008]. A genome-wide role for cohesin in gene expression has also been established, by its association with active enhancers and promoters genome-wide [Seitan et al., 2013]. Finally, in the context of transcriptional regulation, cohesin has been associated with transcriptional termination [Gullerova and Proudfoot, 2008].

Cohesin has been found to co-localise with CTCF during interphase [Parelho et al.,

2008] [Wendt et al., 2008]. In addition to extensive co-localisation throughout the genome, cohesin has been found to specifically interact with CTCF via its Scc3 subunit at the *c-myc* locus. CTCF has been found to be required for recruitment of cohesin to chromatin genome-wide, although a CTCF independent role has also been described. Finally, cohesin has been implicated in imprinted expression through allele-specific co-localisation with CTCF at the *H19/Igf2* and the *Kcnq1ot1* loci [Stedman et al., 2008] [Lin et al., 2011] and a role in allele-specific chromatin structure has been proposed [Nativio et al., 2009].

In combination with CTCF, cohesin has also been implicated genome-wide in chromosomal domain organisation, by the demarcation of loops that define independently regulated domains [Sofueva et al., 2013], functional interactions within pre-exisiting chromosomal compartments [Seitan et al., 2013] and formation of long-range chromosomal interactions at developmentally regulated loci [Hadjur et al., 2009].

1.7 Specific Aims of the Investigation

The work presented here sought to investigate two questions relating to the role of epigenetic processes in two distinct systems, heart and brain.

The genome-wide role of DNA methylation was examined in a model of the mouse developing endocardium. The endocardium examined was isolated from embroid body differentiation culture at the equivalent of embryonic day 9.5 of mouse embryo development. The methylome profile of the endocardium was compared to that of other endothelial tissue from the same model with the aim of identifying distinct epigenetic regulators that drive the phenotypic and morphological differences between the endocardium and the endothelium.

The transcriptomes of the endocardium and endothelium were also examined in order to evaluate the extent to which transcriptional regulation accounts for phenotypic differences between the endocardium and endothelium later in development and to identify differentially regulated genes between the two tissues. Furthermore, motif analysis was used to identify transcription factors that potentially bind and regulate differentially expressed genes.

In addition the transcriptomic and epigenetic datasets generated were combined in order to identify the extent to which epigenetic differences coincide with transcriptomic differences and potentially regulate the latter.

In mouse brain, the extent of genome-wide allele-specific binding of CTCF and cohesin at postnatal day 21 was assessed. The aims of this investigation were to identify the prevalence of allele-specific binding of these two nuclear organisation factors and furthermore identify the extent to which allele-specific binding of these two factors colocalise. The extent to which allele-specific binding of these factors contributes to known imprinted sites was assessed, with the aim of understanding the uniformity of the mechanisms that establish and propagate imprinted gene expression in a genome-wide scale.

1.8 Summary

Epigenetic processes include DNA methylation and histone modifications and contribute to gene regulation, including imprinted gene expression, and development.

The heart is the first functional organ during development and is of critical importance for the subsequent development of the embryo. Multiple lines of evidence suggest that the development of the heart involves epigenetic processes. The endocardium, an endothelial layer lining the internal surface of the heart is a cardiac population of high significance in valvular development and heart remodelling, but remains relatively understudied. The advent of next-generation sequencing methodologies, creates the opportunity to epigenetically and transcriptomically characterise populations of cells such as the endocardium to an unprecedented level. The thesis presented examines the methylome and transcriptome of the endocardial cell population.

CTCF is a Zinc-finger protein primarily known as an insulator, but with a known role in transcriptional regulation. CTCF is known to be involved in parent-of-origin regulation of gene expression in certain loci, but the role of this protein has not been assessed genome-wide in an allele-specific manner. Furthermore cohesin, a protein complex with a known role in sister chromatin cohesion during cell division, is known to co-localise with CTCF during interphase and may have a previously unappreciated role in imprinting. This thesis examines the genome-wide distribution of these two proteins in an allele-specific manner as well as their contribution to known imprinted loci.

Chapter 2

Materials and Methods

2.1 Drop Culture, Differentiation and FAC Sorting of Transgenic ES Cells

The culture and differentiation of transgenic mouse embryonic stem (ES) cells, presented in this section, was carried out in Vanderbuilt University, TN, USA by Mr Kevin Tompkins in the Baldwin Laboratory.

Mouse ES cells previously transfected with the NFATc1-mCherry construct [Misfeldt et al., 2009] were washed twice with phosphate buffered saline (PBS) . Mouse embryonic fibroblasts (MEFs) were depleted by lifting cells with trypsin and re-plating on 0.1% w/v gelatin for 30 minutes. Supernatant from the attached MEFs was transferred to a 15 mL tube and centrifuged at 500 g for 5 minutes. Supernatant was removed and the cells washed twice in differentiation media (DM) [IMDM (Invitrogen, Cat No. 21056), 15% FBS Atlas (Lot No. 80312), 2 mM L-glutamine, Penicilin/streptomycin 1000 U/mL, Transferrin 200 μg/mL, 500 mM L-ascorbic acid, 0.45 mM M-Monothioglycerol]. After the second wash, cells were spun down as above and resuspended in DM at 25,000 cells/mL. Using a 8 channel pipette ten rows of 20 μL were transferred onto the lid of a square Petri dish.

After 2 days of growth, cells were washed from the Petri dish using 3 mL of DM media and transfered to a 15 mL conical vial. Volume was adjusted to 15 mL and EBs were allowed to settle at the bottom on the tube. The supernatant was removed and EBs were resuspended in 10 mL of DM media. The suspension was then transferred to a a 100 mm suspension dish (Fisher Cat. No 430591). The cells were placed on an orbital shaker at 35 revolutions per minute for 48 hours. After 48 hours the cells were plated on 0.1% w/v

gelatin coated 100 mm dishes and media changed after 48 hours.

Plated cells were washed with 10 mL PBS and 5 mL of fresh Accutase were added for 15 to 30 minutes at room temperature on platform shaker. Cells were gently pipetted repeatedly and 5 mL of FACS bufferer were added. The suspension was filtered with a 100 µL sterile cell strainer (Fisher, Cat No: 22363549) for cell clumps and centrifuged at 500 g for 5 minutes. The supernatant was removed and the cells were re-suspended in 10 mL of FACS buffer. Cell count was estimated and cells were centrifuged as above and resuspended at a density of 10^6 per 100 µL in FACS buffer. Anti-CD31 antibody (BD Pharmingen, Cat No: 551262) was added at 1:200 dilution and cells were incubated at 4 °C for one hour with agitation every 5 minutes. Cells were washed 3 times in FACS buffer and re-suspended at a density of 5 10^6 cells/mL in FACS Buffer. Cells were sorted in the Vanderbuild University FACS Core Facility. The sorted cells were centrifuged and the supernatant removed. The cells were snap frozen in liquid Nitrogen and shipped on dry ice to London, UK.

2.2 Next-generation Sequencing

Next-generation sequencing (ng-seq) refers to a set of technologies that allow the sequencing of DNA samples with very high-throughput and low per-base cost. These technologies allow interrogation of the entire genome in a single experiment. Competing technologies are available and the focus here will be on the Illumina®platform as this was utilised in the context of this work. The Illumina®technology is based on millions of reversible dye-terminators reactions performed in parallel on DNA immobilised and locally amplified on a glass surface via bridge PCR (Figure 2.1 B). The Illumina®HiSeq 2000 instruments can generate up to ten billion base-pairs of sequence data per experiment in the course of approximately two weeks. First applications of next-generation sequencing have focused on the sequencing of genomic sequence. However, the technology is now routinely applied to sequencing of the transcriptome, identification of (non-histone) protein binding sites, identification of DNA methylation with base-pair resolution as well as identification of histone modifications.

Utilisation of the technology requires the preparation of a 'DNA library', a solution of DNA in which all the DNA fragments are of relatively uniform size and have identical sequence (adaptors) at their ends, though which they can be manipulated (Figure 2.1 A).

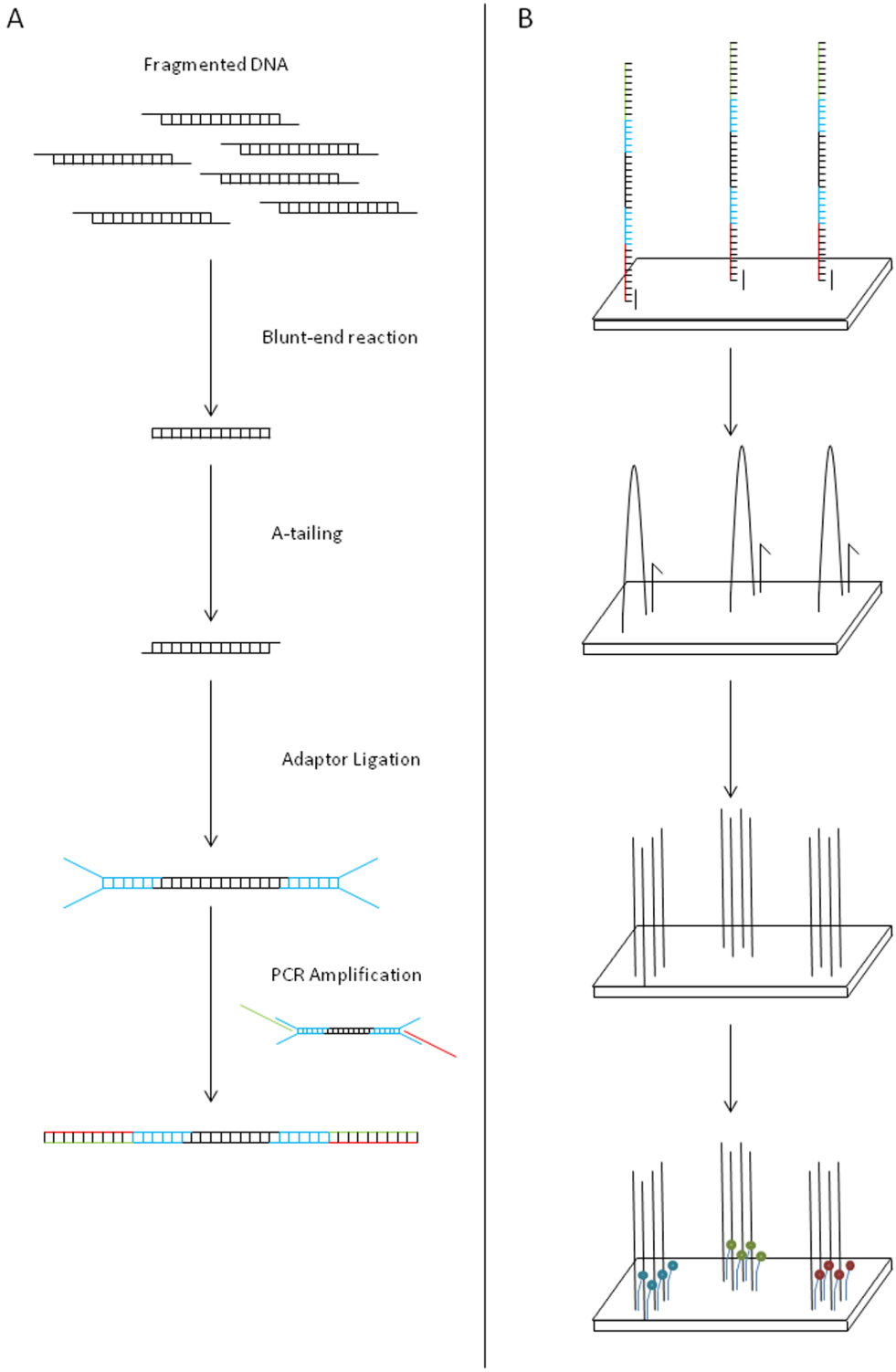


Figure 2.1: A. Outline of library preparation for ng-seq. Overhangs of fragmented DNA are removed and an 'A' overhang added. Adaptors are ligated and all fragments are amplified. B. Outline of sequencing reactions: the library is bound to the flowcell and amplified via bridge PCR. Sequencing is performed using reversible terminator technology.

The technology allows sequencing of fragments present in the library and yields data files containing read sequences (raw reads) along with quality information. The information in the raw reads is then processed bioinformatically to produce useful information.

After quality control, the first step in processing is almost invariably alignment. In the alignment step, the origin of the reads in the reference sequence is identified. The reference sequence is usually the reference genome of the organism in question, although for some applications – notably RNA-seq this may be different. Alignment is performed using an aligner program. The processing of the ng-seq data following alignment differs depending on the type of the experiment and the downstream application. Downstream processing is examined in more detail in the following sections.

It must be noted that sequence generation can be performed in two distinct modes, single-end sequencing and paired-end sequencing. In single-end sequencing, every DNA fragment is sequenced from one end generating a sequence that can be individually mapped to the genomic reference. In paired-end sequencing, each DNA fragment is sequenced from both ends generating two linked read sequences that are expected to map in close proximity on the genomic sequence. In addition to increasing available sequence, paired-end sequencing allows for better mapping of raw reads by constraining the mapping of read pairs to location where both read align condordantly.

2.2.1 Histone and DNA Binding Factor ChIP-seq

Chromatin Immunoprecipitation followed by next generation sequencing (ChIP-seq) allows the identification of binding sites of both histone and non-histone proteins across the genome [Valouev et al., 2008]. Furthermore, it can be used to interrogate histone modifications across the genome. The methodology is based on chromatin immunoprecipitation followed by ng-seq. In ChIP protocols, DNA quantification can be performed via quantitative PCR (qPCR) in a locus-specific fashion. In ChIP-seq, the precipitation is instead coupled to ng-seq by using the output of the precipitation as the input for the preparation of a ng-seq library.

ChIP is based on the concept of specifically selecting DNA binding factors from the chromatin solution via a specific antibody (Ab) and simultaneously isolating the DNA to which they are specifically bound (Figure 2.2). For this reason, ChIP protocols are frequently preceded by a cross-linking step to reinforce DNA-protein interactions via covalent links, although its importance is diminished – and the step is sometimes omitted

– when expected protein-DNA interactions are strong, as is the case for histone ChIP. Furthermore, to allow for the specific isolation of the DNA binding to factors of interest, the chromatin suspension is sonicated to fragment the DNA and prevent pull-down of distant DNA fragments via the DNA sugar-phosphate backbone. After the precipitation is performed, the cross-links are reversed and the proteins are digested away allowing for the identification of the DNA fragments to which the proteins of interest were bound via ng-seq.

ChIP-seq Data Processing

The aim of ChIP-seq data processing is to identify distinct genomic locations, 'peaks', where enrichment has occurred during immuno-precipitation. The first step in ChIP-seq data processing is the mapping of reads to the genome, in a similar way to that for other ng-seq experiments. Following alignment, identification of peaks and assignment of a significance score is performed [Pepke et al., 2009].

Identification of peaks is a complex process owing to the fact that the distribution of reads that arise from the non-enrichment based processes is not random, but displays highly complex biases. For this reason an input sample is usually sequenced, allowing identification of locations of high read density where enrichment has not occurred [Pepke et al., 2009].

Due to the fact that fragmentation of DNA occurs while the protein of interest remains bound, the distribution of reads surrounding any peak display a characteristic bimodal distribution (Figure 2.3) [Anders and Huber, 2010]. The distance difference between the reads that align to the two sides of the peaks (the 'peak shift') is related to the insert size in the case of paired-end sequencing, but as it cannot be accurately predicted from it, it is empirically estimated. After identification of the peak shift, reads are shifted so as to produce distinct peaks that can be computationally identified (Figure 2.3).

Identification of peaks is performed after peak shifting with specialised computer programs, referred to as 'peak callers'. Over 30 different peak callers, such as Useq [Nix et al., 2008] and MACS [Zhang et al., 2008], have been developed making selection of the appropriate algorithm difficult [Wilbanks and Facciotti, 2010]. The approaches of these programs vary dramatically, from simple count based sliding window approaches to specific search for strand-specific read patterns indicative of binding [Wilbanks and Facciotti, 2010].

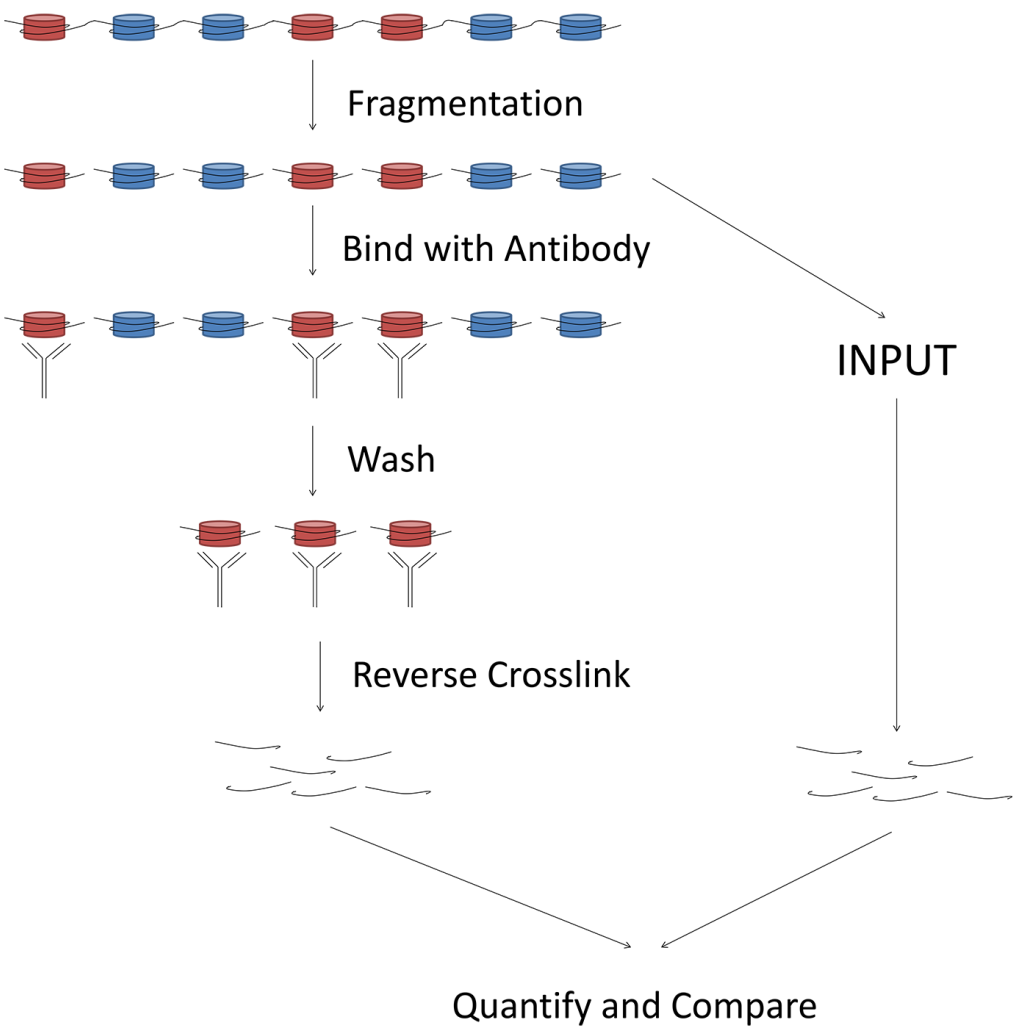


Figure 2.2: Overview of ChIP protocol for histone modifications. Chromatin is fragmented via mechanical or enzymatic means. Nucleosomes with modifications of interest are conjugated to an immobilised specific antibody. Non-specific interactions are abolished via washing the sample. The DNA bound to the nucleosomes of interest is eluted via reversing crosslinks and quantified against the general chromatin background. Quantification can be performed via qPCR or in the case of ChIP-seq via ng-seq.

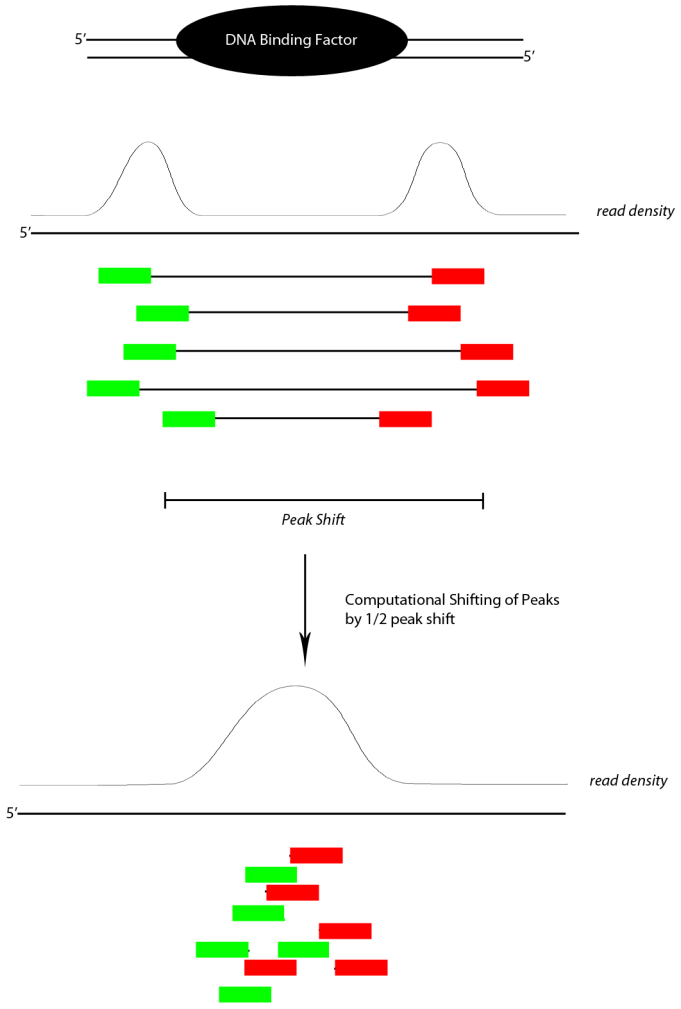


Figure 2.3: Overview of computational correction of the peakshift in ChIP-seq read data. ChIP-seq reads align on either side of the binding sites of the protein of interest. Identification of this offset allows for the correction of the read position via computational shifting of the peaks and yields better defined binding events.

Allele-specific ChIP-seq

ChIP-seq is generally performed in such a manner that differential binding of a protein, or the presence of a histone modification between the two parental alleles can not be distinguished. If however, the tissue on which ChIP-seq is performed on originates from the offspring of a cross between two divergent strains of animals, polymorphisms – such as single nucleotide polymorphisms (SNPs) – between the two parental alleles can be used to identify the allelic origin of some of the the next-generation sequencing reads. The strains of animals used for the cross need to be sufficiently divergent so that a large proportion of reads overlap a polymorphism. In this manner, the origin of each of the peaks identified by a ChIP-seq experiment can be traced to specific parental alleles. In order to discriminate parent-of-origin effects from sequence-specific effects, reciprocal crosses are employed, whereby the dam and the sire strains are exchanged and the experiment is performed on animals obtained from both crosses.

2.2.2 Whole-genome Bisulphite Sequencing

Whole genome bisulphite sequencing (WGBS) or BS-seq refers to next-generation sequencing coupled with bisulphite conversion of DNA [Bock, 2012] [Wu et al., 2011b]. As detailed in Section 1.1.3, DNA can be methylated at the 5 carbon position of C. Interrogation of mC can be performed by means of bisulphite conversion followed by sequencing (Figure 2.4). Bisulphite conversion can be used to identify mC because incubating DNA with sodium bisulphite exclusively converts unmethylated C to uracil (U), whereas mC is unaffected. U is then converted to thymine (T) via the action of a non-uracil sensitive DNA polymerase. After bisulphite treatment, the DNA can be sequenced via Sanger sequencing or in the case of WGBS via next-generation sequencing. By comparison to the original sequence mCs can be identified. In this way, BS-seq allows genome-wide identification of methylation status at a single base-pair resolution.

Processing of Whole-genome Bisulphite Sequencing Data

Bioinformatic processing of bisulphite next-generation sequencing data is more challenging than that of data obtained from other next-generation sequencing experiments because the sequence data can no longer be directly aligned to the genomic sequence [Krueger and Andrews, 2011]. Furthermore, difficulties in aligning are compounded by the fact that the majority of the sequence generated is composed of only three bases (A, T and

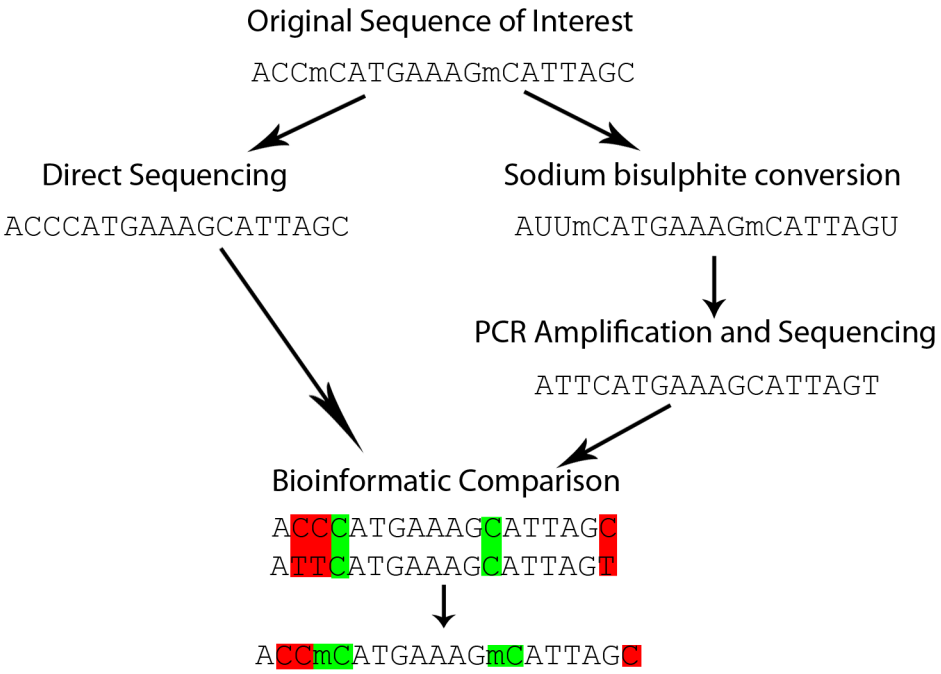


Figure 2.4: Outline of determination of methylation status of cytosine via sodium bisulphite conversion on a sample sequence. The DNA is sequenced twice, once via conventional sequencing and once after treatment with sodium bisuphite and amplification. Unmethylated Cs are read as Ts after conversion, whereas mC are not changed, allowing bioinformatic determination of the original modifications.

G), as opposed to four, making unique alignment of reads much more challenging. For this reason, long paired-end (in excess of 100 bp) reads are employed in WGBS. This aids alignment by providing longer sequence, improving confidence of the alignment and discriminating between competing alignment possibilities.

Alignment of BS-seq data is accomplished by aligning *in silico* fully converted reads (i.e. reads for which even methylated Cs are converted to Ts) to a fully *in silico* converted genome [Krueger and Andrews, 2011]. This ensures that the methylation status of different loci does not affect the probability of alignment, ameliorating any potential bias. The alignment is further complicated by the fact that *in silico* conversion of DNA gives rise to four possible genomic sequences to align to because the final sequence depends on which strand the bisulphite conversion is applied to. All possible sequences are prepared and alignment is carried out against all four in the case of non-directional libraries, or only two when libraries are prepared in a way that preserves directionality.

2.2.3 RNA-seq and mRNA-seq

Sequencing of RNA species derived from a biological sample through the use of next-generation sequencing is referred to as RNA-seq [Wang et al., 2009]. RNA-seq has some significant differences, in the experimental protocol and data processing steps that are the result of the sequencing of RNA as opposed to DNA species. Unlike traditional sequencing libraries, RNA-seq utilises an initial reverse transcriptase reaction to convert RNA to copy DNA (cDNA) before preparation of the sequencing library in a manner similar to that previously outlined.

In comparison to other methodologies that can sequence and quantify RNA, such as microarrays, RNA-seq has the advantage that it can interrogate the transcriptome without being limited to known transcribed loci, allowing the discovery of novel tissue-specific transcripts and not introducing a bias towards previously examined loci [Mortazavi et al., 2008]. Furthermore RNA-seq can accurately quantify transcript levels with a greater dynamic range and better accuracy [Fu et al., 2009].

A major drawback of RNA-seq is that due to the high levels of ribosomal RNA (rRNA) transcripts being present in cells, a large percentage of the sequenced species originate from rRNA loci. This results in the majority of the generated sequence consisting of multiple copies of the rRNA species that are not otherwise informative of differences in cellular function.

The mRNA-seq protocol attempts to circumvent this problem by allowing only RNA species with a poly(A) signal to be incorporated in the library. This is accomplished by specific selection of RNA species containing the poly(A) signal by means of immobilised poly(T) sequences. Other methodologies for the removal of rRNA are available, such as specific removal of ribosomal RNA. Specific removal of the ribosomal RNA has the distinct advantage in comparison to mRNA-seq that it can sequence non-poly-adenylated transcripts that are not rRNA and are functional, such as microRNA species.

RNA-seq allows the accurate quantification of RNA species. It is however known that the reverse transcriptase reaction and the subsequent PCR amplification can introduce artefacts. More recent methodologies that allow detection and elimination of such biases through the use of unique molecular barcodes have been developed [Islam et al., 2014].

RNA-seq Data Processing

As outlined above, RNA-seq data processing differs significantly compared to the processing of other ng-seq data. Unlike other protocols, RNA-seq data is usually not aligned to the genomic reference, but rather to a *in silico* generated transcriptome, that takes into account splicing events and allows mapping across exon junctions. Optionally mRNA-seq data can be aligned to the genome in an attempt to identify novel splice sites and expression of non-annotated genomic loci [Trapnell et al., 2010].

Following mapping, the reads aligning to each transcript are quantified and compared between conditions to identify differentially regulated genes or transcripts. Quantification can be performed by obtaining a simple count [Anders and Huber, 2010] of the number of reads that align to each transcript or by calculation of a metric such as FPKM (Fragments per Kilobase per Million reads) that normalises the reads counts for the length of each transcript and for the library size [Trapnell et al., 2013].

The advantages of either approach are not clear. The main argument for the FPKM approach is that it takes into account the size of the transcript in question and the library size. However, consideration of the overall transcript length is not important for identifying the differential expression of the same transcript between two conditions. Comparison of expression level of transcripts within the same sample, where an absolute measure of expression would be of value, is not however recommended as sequencing efficiency of different transcripts can vary widely. In contrast count based methods, such as DESseq, rely on a statistical framework that is more robust and can still take into

account the overall library size.

2.3 Endocardial and Endothelial Cell Methylome Analysis

2.3.1 WGBS Library Preparation and Optimisation

The protocol development, optimisation and initial replicate was performed by the author. The second replicate was performed by Mr Samuele Maria Amante.

The protocol for whole genome bisulphite sequencing (WGBS) library generation was developed from the Reduced Representation Bisulphite Sequencing (RRBS) protocol published by Meissner and colleagues [Meissner et al., 2005]. In particular, the protocol was modified and enzyme based fragmentation was replaced by sonication as described below.

DNA for WGBS was isolated by Mr Kevin Tompkins in the Baldwin Laboratory and shipped on dry ice. Approximately 100 ng of DNA were diluted with Tris-EDTA Buffer in a 1.5 mL LoBind tube (Axygen, Cat No. MCT-150-L-C) to a total volume of 120 µL. The DNA was carefully loaded in a Covaris sonication tube (Covaris, microTUBE AFA Fiber pre-split 6x16mm, Part No.: 520045, Lot No.: 001958) and sonicated (duty cycle 10%, intensity 4%, cycles per burst 200, time 6 cycles of 70 s each, frequency sweep mode, temperature 4°C to 7°C). The fragmented DNA was then transferred to a clean 1.5 mL LoBind tube. The DNA was then concentrated using the QIAGEN MinElute ®PCR Purification Kit (Cat. No: 28004, Lot: 148027401) and eluted in 44 µL of Elution Buffer.

Next-generation sequencing libraries were prepared using the NEBNext®ChIP-Seq Library Prep Master Mix Set for Illumina®(Cat. No. E6240S, Lot No: 0131201) according to manufacturers' instructions until up to and including the dA-tailing step. Cleanup after End Repair, dA-tailing and Adaptor Ligation was performed using the QIAGEN MinElute ®PCR Purification Kit (Cat. No: 28004, Lot: 148027401). For replicate 1, custom pre-hybridised methylated paired-end (PE) adaptors from Sigma were used during adaptor ligation, whereas for replicate 2 NEBNext®Multiplex Oligos for Illumina®(Methylated Adaptor, Index Primers Set 1) (Cat. No E7535S) adaptors were used.

After adaptor ligation, bisulphite conversion was performed using the Zymo Research, EZ DNA Methylation Kit (Cat No: D5001, Lot No.: ZRC176186) in a LoBind tube. The incubation protocol for methylation arrays was used (15 cycles of 1 hour 50 °C incubation with 95 °C temperature spike for 30 seconds and hold at 4 °C for at least 10 minutes). Desulphunation was performed on the column as per manufacturers' instructions and the

converted DNA was eluted in 10 µL of molecular biology grade water.

DNA was first amplified with Pfu Turbo Cx Hotstart DNA polymerase Agilent, (Cat. No. 600410). The reaction was prepared as shown in Table 2.1. Custom HPLC purified PE primers from Sigma were used. The solution was mixed thoroughly by pipetting and split into five 10 µL aliquots and placed in a thermocycler (MJ Research PTC-200). The program shown in Table 2.2 was executed.

Table 2.1: Reaction composition of WGBS library initial amplification with Pfu DNA polymerase.

Bisulphite converted DNA	5 µL
10X turbo buffer	5 µL
PE primer 1 (5 µM)	2 µL
PE primer 2 (5 µM)	2 µL
dNTPs 10mM each	1 µL
Pfu Cx polymerase Hotstart (Agilent, Cat. No.: 600410)	1 µL
Sterile Nuclease Free H_2O	34 µL
Total	50 µL

Table 2.2: Reaction temperature cycle of WGBS library initial PCR amplification with Pfu DNA polymerase.

1	Initial denaturation	95 °C	2 minute
2	Denaturation	95 °C	30 seconds
3	Anneling	65 °C	30 seconds
4	Extension	72 °C	45 seconds
5	Repeat from 2	7 times	
6	Final Extension	72 °C	10 minutes
7	Temperature Hold	4 °C	indefinitely

The DNA from the five reactions was pooled and purified using the Qiagen Minelute PCR Purification Kit (Cat. No.: 28004) and was amplified again using Platinum Pfx Turbo DNA polymerase. The Pfu PCR reaction composition is shown in Table 2.3.

The solution was mixed thoroughly by pipetting and split into five 10 µL aliquots and placed in a thermocycler. The program shown in Table 2.4 was executed.

The PCR mix was electrophorised on a 3% w/v low melting point NUSIEVE ®GTG ®agarose (Cat. No: 50080, Lot: AG2475) gel for 1 hour at 85 V and a band from 300 bp to 700 bp was excised. DNA was extracted using the Qiagen Minelute Gel Extraction Kit (Cat No: 28606) and following manufacturers' instructions. The DNA was not heated after QC buffer addition but vigorously vortexed for at least 5 minutes and until no gel was visible. This modification of the protocol was introduced because heating of the

Table 2.3: Reaction composition of WGBS library second amplification with Platinum Pfx DNA polymerase.

DNA from Pfu Reaction	10 µL
10x Pfx Buffer	5 µL
$MgSO_4$ (50 mM)	2 µL
dNTPs 10 mM each	2 µL
PE Primer 1 (5 µM)	2.5 µL
PE Primer 2 (5 µM)	2.5 µL
Platinum Pfx polymerase (Invitrogen, Cat No: 11708-013)	0.8 µL
Sterile Nuclease Free H_2O	25.2 µL
Total	50 µL

Table 2.4: Reaction temperature cycle of WGBS library second PCR amplification with Platinum Pfx DNA polymerase.

1	Initial denaturation	94 °C	2 minutes
2	Denaturation	94 °C	20 seconds
3	Anneling	65 °C	30 seconds
4	Extension	68 °C	30 seconds
5	Repeat from 2	11 times	
6	Final Extension	68 °C	5 minutes
7	Temperature Hold	10 °C	indefinitely

sample at this stage has been found to introduce a GC bias [Quail et al., 2008]. The DNA was eluted in 10 µL and quantified with the quBit instrument (Section 2.4) and its size distribution examined with the Agilent Bioanalyser or Agilent Tapestation (Section 2.5).

2.3.2 WGBS Library Sequencing

Cluster generation and sequencing of the WGBS libraries was performed in the BRC Sequencing Facility on a HighSeq 2000 Illumina® Sequencer. A low amount phiX control DNA was added by the sequencing facility for quality control purposes.

2.3.3 Bioinformatic Processing of WGBS Data

Basecalling of the raw read data was performed with Casava (version 1.8.0) on the BRC HPC computing cluster, the following parameters were specified in addition to the standard parameters specifying data location paths.

```
--use-bases-mask Y100,N7,Y100
--no-eamss
--positions-format _pos.txt
--force
```

```
--fastq-cluster-count 500000000
--ignore-missing-stats
--ignore-missing-bcl
--ignore-missing-control
```

The quality of the read data was assessed using the FastQC program (version 0.10.0). Sequences identified as of low quality by the sequencer software were removed using the `fastq_illumina_filter` utility. Adapter sequences were removed using `TrimGalore` (version 0.2.5).

The phiX genome was obtained in FASTA format from RefSeq (Sequence ID: NC_001422.1). The Bowtie index of the phiX genome was built with `bowtie-build`. Reads were aligned to the phiX reference genome using the Bowtie 1 aligner (version 0.12.8) [Langmead et al., 2009] with the `--un` flag to keep unaligned reads and save them for later processing. Reads that successfully aligned to the phiX genome were discarded.

The remaining (unaligned) reads were aligned to the bisulphite converted mm9 version of the mouse genome using Bismark (version 0.14.2) [Krueger and Andrews, 2011] and Bowtie 1 (version 1.1.1) [Langmead et al., 2009] as the underlying aligner. Methylation extraction was performed using the `bismark_methylation_extractor` tool from the Bismark package [Krueger and Andrews, 2011].

Methylation data from the `bismark_methylation_extractor` tool were summarised on a per genomic position basis using a custom Perl script. The script counted instances of methylated and unmethylated occurrences of Cs in the data for each genomic position independently (Appendix Section C.2).

Summarisation to CGIs and testing for differential methylation was performed using custom scripts developed in the context of this project. Whole-genome analysis was performed using the `bsseq R` package (version 3.1) [Hansen et al., 2012] . Mapping identified differentially methylated region and CGIs to genomic element classes was performed with the CEAS tool [Shin et al., 2009].

Examination of the overlap of the identified DMRs with cardiac histone marks and DNAse hypersensitivity sites was performed using custom `R` and UNIX shell scripts (Section C.3). Permutation testing was performed by generating permuted datasets with the bedtools suite `shuffle` command and counting the number of overlaps between each mark and the identified peaks. 1,000 permutations were performed for each mark.

2.4 DNA Sample and Library Quantification

The Qubit fluorometer instrument with the Qubit dsDNA HS Assay Kit (Cat No: Q32854, Lot: 1429806) was used for the quantification of all libraries prior to final dilution and sequencing. Samples and reagents were allowed to reach room temperature for at least 30 minutes prior to quantification. Samples were prepared according to manufacturers' instructions. Specifically, 200 μL of working solution were prepared per sample analysed by mixing DNA HS Reagent with Qubit buffer in 1:200 ratio. 1 μL of sample, in some cases pre-diluted 1:10 as described, was diluted to a total volume of 200 μL with working solution, mixed and incubated for 3 minutes. Similar preparation was performed for standards, but 10 μL or each standard were diluted in a total volume of 200 μL of working solution. Instrument calibration was performed prior to every set of measurements.

2.5 Next-generation Sequencing Library Size Estimation

Library size estimation was performed with either the Agilent Bioanalyser or the Agilent TapeStation instruments in the BRC Genome Sequencing Facility.

The Agilent Bioanalyser with the DNA 1000 Kit was used for identification of the fragment size distribution of some DNA samples. Manufacturers' instructions were followed for chip sample preparation. Specifically, gel-dye mix was prepared by allowing the reagents to reach room temperature for 30 minutes, the dye concentrate was vortexed and spun down and 25 μL of dye were added to the DNA matrix vial. The tube was vortexed and transfered into a spin filter and spun in a microcentrifuge for 15 minutes at 2240 g. In some cases previously prepared gel-dye mix, provided by the Genomic Sequencing facility, was used. 9 μL of the prepared gel-dye were added to the G well of a new DNA chip and the chip was primed using the priming station for 60 seconds and released. After 5 seconds the plunger was returned to the 1 mL position and the chip was removed from the priming station. 9 μL of the gel-dye mix were pipetted into the appropriate wells. 5 μL of DNA marker was pipetted into each of the sample wells and 1 μL of marker ladder was pipetted into the ladder well. 1 μL of sample or deionised water was pipetted into each sample well.

The Agilent TapeStation instrument was alternatively used with the D1K HS Kit to identify fragment size distribution of DNA samples, as noted in the text. Manufacturers' instructions were followed for sample preparation. 2 μL of DNA sample were mixed with

1 µL of marker and loaded into the Tapestation instrument.

2.6 qPCR Quantification of Library Concentration

qPCR quantification of pooled libraries was performed to ensure correct balancing of the individual component libraries. Quantification was performed using the KAPA Biosystems Illumina Quantification Kit ABI Prism ®qPCR Mix (Cat No. KK4835, Lot No: Z48350000271) according to manufacturers' instructions.

A 1:1000 dilution of each library was prepared and a qPCR plate with technical triplicates of the reaction shown in Table 2.5 was prepared along with one reaction for each of the standard concentrations. The qPCR protocol shown in Table 2.6 was executed on the qPCR instrument. Library concentration was calculated against the 452 bp standard as outlined in the results section.

Table 2.5: KAPA SYBR qPCR reaction composition.

KAPA SYBR ®FAST qPCR Master Mix containing Primer Premix	12 µL
PCR-grade water	4 µL
Diluted library DNA or DNA standard (1 - 6)	4 µL
Total	20 µL

Table 2.6: KAPA SYBR qPCR reaction temperature cycle.

1	Initial activation denaturation	95 °C	5 minutes
2	Denaturation	95 °C	30 seconds
3	Annealing extension data acquisition	45 °C	45 seconds
4	Repeat from step 2, 34 times		

2.7 Endocardial and Endothelial Cell Transcriptome Analysis

2.7.1 RNA Extraction Protocol

RNA was extracted from FAC sorted endocardial and endothelial cells using the QIAGEN RNeasy®Mini Kit (Cat No: 74104, Lot No: 145038477). All kit reagents and columns were allowed to cool for at least one hour on ice prior to the RNA isolation.

Frozen shipped cells were thawed by addition of ice cold PBS and incubation on ice

for 5 minutes. Cells were isolated by centrifugation at 1600 g for 5 minutes and aspiration of the supernatant. Cells were lysed by addition of 350 µL of RLT buffer and incubation on a Thermoshaker at 1400 rpm for 3 minutes at 4 °C. The lysate was centrifuged for 3 minutes at maximum speed and supernatant was moved to a clean tube. 350 µL of 70% v/v ethanol was added to the lysate and mixed by pipetting 10 times. The sample was transferred to a pre-chilled RNease Mini spin column placed in a collection tube and spun for 15 seconds at maximum speed. The flow-through was discarded and the column was washed successively with 700 µL of Buffer RW1 and 500 µL of Buffer RPE and separated by centrifugation for 15 seconds at maximum speed. Finally 500 µL of Buffer RPE was added and the column was centrifuged for 2 minutes at maximum speed. The column was transferred to a new collection tube and centrifuged at full speed for 1 minute. The column was placed in a new 1.5 mL collection tube and 30 µL of RNA-free water were added to the membrane. The column was incubated at room temperature for 1 minute and spun for 1 minute at maximum speed to elute the RNA.

2.7.2 RNA Quantification and Quality Assessment

RNA quantity was assessed using the QuBit®RNA HS Assay (Molecular Probes, Cat. No. Q32852, Lot: 1416120). All reagents were allowed to reach room temperature for at least 30 minutes prior to quantification and 200 µL of working solution were prepared for each sample and standard by mixing RNA reagent and RNA Buffer in 1:200 ratio. For each standard 10 µL of standard were mixed with 190 µL of working solution and for each sample 1 µL of sample was mixed with 199 µL of working solution. All samples and standards were mixed and incubated at room temperature for 2 minutes prior to quantification.

RNA quality was assessed using the TapeStation R6K Tape and Reagents (Agilent, Cat No. 5067-5367 and 5067-5368). 4 µL of Sample Buffer were mixed with 1 µL of sample and the samples were incubated at 72 °C for 3 minutes before being placed on ice for 2 minutes. The samples were centrifuged before being analysed with the Tapestation 2200 instrument (Agilent, Cat No: G2964AA).

2.7.3 RNA Extraction Optimisation

Prior to extracting the RNA from endocardial and endothelial cells, the protocol was optimised to ensure that an adequate quantity and quality of RNA could be extracted.

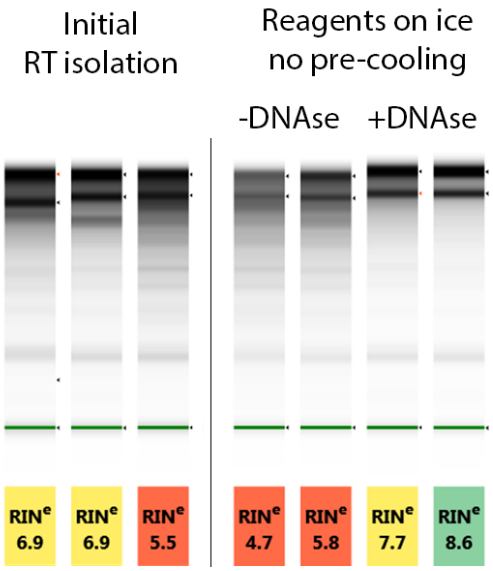


Figure 2.5: Gel representation of Tapestation data for RNA used in RNA optimisation trial. Initial isolation (left) was performed in triplicate and displayed quality of DNA lower than the minimum required (8.0) for library preparation. Repetition with cooled reagents showed improvement, that also correlated with the amount of time the reagents had remained on ice, samples with DNAse were prepared after samples without DNAse.

Optimisation was performed on NIH/3T3 aliquots of 100,000 cells each, provided by Dr Mike Cowley and Mr Samuel Amante. Separate RNA preparations were uniquely identified with letter codes followed by "_trial" as shown in Tables 2.7 and 2.8.

Initial triplicate isolations at room temperature (RT) yielded RNA of low quality as assessed on the Tapestation platform (Figure 2.5). The isolation was repeated in duplicate with and without a DNAse treatment step. Surprisingly, the DNAse treatment improved the RNA quality (Welch Two Sample T-test, p-value <0.032) (Figure 2.5). This was attributed to the fact that the preparation including the DNAse was performed after the incubation without DNAse and reagents and all solutions had remained on ice for a longer period of time. In subsequent preparations all solutions were pre-chilled on ice for a minimum of one hour and the RIN quality score was consistently above 8.0 without DNAse treatment (see Figure 3.20).

The quantity of extracted RNA was assessed on a new RNA preparation using the quBit instrument to confirm that sufficient material could be isolated to meet the minimum requirements (0.1 µg) for the library preparation protocol (Table 2.7).

The stability of purified RNA at 4 °C and at room temperature was also assessed to gain an understanding to the care required when handling RNA prior to cDNA synthesis. Results are presented in Table 2.8. RNA was incubated at 4 °C or at RT and the RIN

Table 2.7: RNA quantification of trial RNA preparation.

Sample ID	Concentration (μg/mL)			Volume (mL)	Total RNA (μg)
	1st	2nd	Mean		
E_trial	24.3	24.2	24.3	50	2.15

Table 2.8: RNA stability assessment of trial RNA preparation. RNA is incubated at the defined temperature for the time indicated and the RIN is measured. RNA was stable for several hours both at 4°C and Room Temperature (RT).

Time(hours)	0	5	24
Temperature (°C)		4	RT
Sample Id			
B_trial	5.8	5.7	N/A
D_trial	8.6	8.5	8.4

number was measured at the end of the incubation period. RNA is essentially stable at RT for over 24 hours following purification, suggesting that degradation occurs during extraction and further supporting the importance of pre-chilled reagents.

2.7.4 RNA-seq Library Preparation and Sequencing

Snap frozen cells, provided by the Baldwin lab, were thawed by addition of 0.5 mL ice-cold PBS and RNA was extracted using the QIAGEN RNeasy Mini Kit as described in Section 2.7.1 by the author. All reagents of the QIAGEN RNeasy Mini Kit and all tubes used were pre-chilled on ice for a minimum of one hour. RNA samples were kept on ice at all times. RNA quality was assessed as described in Section 2.7.2 and the quality of the RNA was found to be equal or above 8.0 for all samples that were used for sequencing.

RNA-seq libraries were prepared immediately following RNA purification, using the Illumina®Stranded mRNA LT kit (Part Number: RS-122-2101, LOT: 401884) following the published low throughput protocol from the TruSeq Stranded mRNA Sample Preparation Guide (Part Number 15031047, Rev. D, September 2012). Fragmentation time was modified to 1 minute as per Table 21 of Appendix A of the Illumina sample preparation guide to one minute, to allow for longer insert sizes compatible with the 100 bp paired-end sequencing protocol.

Cluster generation and sequencing was performed by the BRC Sequencing Facility. Sequencing was performed in paired-end mode with 100 to 101 cycles per read with indexing.

2.7.5 RNA-seq Data Basecalling, Quality Control and Alignment

Basecalling was performed on the BRC HPC Cluster with Casava (version 1.8.0). Configuration of the casava parameters was performed using the `configureBclToFastq.pl` script. The following non-standard parameters were passed to the configuration script

```
--use-bases-mask Y100,I6N3,Y100
--no-eamss
--input-dir
--sample-sheet
--fastq-cluster-count 500000000
```

furthermore the input directory and sample sheet were specified with the options respectively.

Quality control was performed with FastQC (version 0.10.1) [Andrews, 2014]. Reads were trimmed using the `fastx_trimmer` program from the fastx toolkit [Pearson et al., 1997] with the parameters `-Q33 -f 21`, to handle quality scores correctly and trim the first 21 bp of all reads, respectively.

The reference transcriptome was built once prior to alignment of individual samples with the `-p 8 -G [reference-gtf] --transcriptome-index` options on the mm9 version of the mouse genome. Alignment was performed with Tophat 2 [Kim et al., 2013] (version 2.0.10) and Bowtie 2 [Langmead and Salzberg, 2012] (version 2.1.0) with the following command line arguments

```
-p 8 --transcriptome-index [path-to-index]
--library-type fr-secondstrand
```

against the reference transcriptome. Reads that did not align to the transcriptome were aligned to the genome.

Duplicate identification and removal was performed with the Picard Toolkit version 1.105 [Li et al., 2009] using the `REMOVE_DUPLICATES=TRUE` flag.

Annotation assembly was performed using cufflinks (version 2.1.1) [Trapnell et al., 2013] for each sample individually starting from the aligned de-duplicated reads. The reference UCSC annotation was provided as a reference with the `--GTF-guide` option. Individual annotations were merged using cuffmerge from version of 2.1.1 of cufflinks. The reference annotation and genomic sequence were provided using the `--ref-gtf` and `--ref-sequence` options.

2.7.6 Identification of Differential Expression

Differential expression was identified with Cuffdiff (version 2.1.1) [Trapnell et al., 2013]. The program was run with the following command line parameters

```
-v -o output --labels EC,ET
--num-threads 8
--library-type fr-secondstrand input/merged.gtf
[Endocardial-bam-files] [Endothelial-bam-files]
```

The `R` package `CummeRbund` [Goff et al., 2012] was used for producing some of the plots of the data.

2.7.7 GO Term Analysis of Differentially Expressed Genes

Gene Ontology overrepresentation analysis of the differentially expressed genes was performed using GO-Elite (version 1.2 beta) [Zambon et al., 2012]. The pruned output results were used for subsequent analysis. Results were filtered in accordance with the recommended settings in Microsoft Excel. Specifically the number of genes changed was required to be greater than two, Z-score was required to be greater than 1.96 and adjusted p-value was required to be less than 0.05.

2.7.8 Identification of Differentially Regulated Transcription Factors

Genes identified as differentially regulated in the transcriptome analysis and annotated with GO term "sequence-specific DNA binding transcription factor activity" (GO:0003700) were identified using a custom SQL script. A copy of the `assocdb` version of the GO term database was retrieved and installed locally from http://www.geneontology.org/ (date of retrieval: 15/02/2014). The database was queried using the mysql script in Section C.4 of the Appendix.

The list of differentially regulated transcription factors was further expanded by merging the results of the above analysis with alternative methods of transcription factor identification. In particular, the intersection of the list of differentially regulated genes with two lists of known transcription factors were independently produced [Kanamori et al., 2004] [Zhang et al., 2012] and merged. Intersection of the lists was performed with custom UNIX shell commands.

2.7.9 Identification of TF Binding Distribution near TSSs

In order to define a search interval for the motif analysis presented in the following section, the distribution of known transcription factor binding peaks was examined. An `R` script (Section C.5) was developed to retrieve ENCODE TF ChIP-seq data from the UCSC browser database and map them to the nearest TSS site, within a search interval of 20 kb, utilising the `ChIPseeker` `R` package [Yu, 2014]. A 20 kb cutoff was selected after trial and error suggested that all the examined TFs bind near baseline levels well prior to this cutoff.

Initially all TF tracks from the ENCODE project [The ENCODE Project Consortium, 2012] were processed, however some non-canonical TFs were identified in this dataset, such as CTCF and cohesin. All proteins for which ChIP data was available were subsequently checked in the NCBI gene database for definite evidence of transcription factor function and only those for which such evidence could be found were subsequently used. The examined transcription factors comprised *E2f4*, *Max*, *Nrsf*, *Tcf3*, *Usf1*, *Gata1a*, *Tal1*, *Gata2*, *Ets1*, *Mxi1*, *Nrf2*, *Gcn5* and *Usf2*.

2.7.10 Transcription Start Site Motif Analysis

Sequences for TSS motif analysis were obtained using a custom R script (Section C.6). Overlapping genomic windows of 1 kb each with a 0.5 kb overlap were obtained for the region extending 5 kb downstream to 5 kb upstream of each TSS of interest preserving directionality of the TSS. The positive dataset consisted of the TSS regions of significantly upregulated genes, whereas the negative dataset of genes that were not identified as differentially regulated, exhibited a log2(fold change) smaller than 0.02, were marked as having sufficient data by cuffdiff and had a mean FPKM greater than 10, these genes were selected using a custom `awk` script from the `cuffdiff` output.

Motif analysis was performed with the online DREME tool (version 4.10.0 patch 4) [Bailey et al., 2009]. Identification of motifs was performed with the online version of the TOMTOM tool [Bailey et al., 2009] against the 'JASPAR Vertebrates and UniPROBE Mouse' database.

The distribution of the identified motifs was examined with the FIMO tool of the MEME Suite. The output of the tool was processed with a custom `R` script to generate the distribution of the individual motifs.

2.8 Identification of Allele-specific CTCF Binding Sites in the Mouse Brain

2.8.1 CTCF and Cohesin ChIP-seq

This work was performed by Dr Adam Prickett prior to the commencement of the work presented here.

CTCF and Cohesin ChIP-seq was performed on postnatal day 21 (P21) mouse brain on F_1 offspring of crosses between C57Bl6 (Bl6) females and *Mus musculus castaneus* (*castaneus*) males (BxC) and *vice versa* (CxB).

Chromatin Isolation

This work was performed by Dr Adam Prickett prior to the commencement of the work presented here.

Banked P21 mouse brains were homogenised in 1 mL of PBS pH 8.0 in the presence of Protease Cocktail Inhibitors (Roche, Catalogue Number: 04693132001). Nuclei were separated by centrifugation at 5000 rpm for 5 minutes. All subsequent centrifugations were at 5000 rpm for 3 minutes and resuspension was in 1 mL of PBS. Nuclei were washed 3 times and cross-linked in 1 mL of 5 mM DTBP (Pierce, Catalogue Number: 20665) on ice for 30 minutes. The nuclei were washed twice in PBS, followed by a wash in 0.1 M Tris-HCl pH 8.0, 1.5 M NaCl and two washes with PBS. Nuclei were cross-linked in 1% formaldehyde in PBS for 10 minutes at 37 °C and washed three times in PBS. Lysis was performed in 50 mM Tris-Hcl pH 8.0, 1% w/v SDS, 10 mM EDTA supplemented with 0.1 mM PMSF.

DNA quantification was performed with the Nanodrop instrument (Thermo Scientific).

Chromatin Sonication

This work was performed by Dr Adam Prickett prior to the commencement of the work presented here.

Chromatin sonication was performed using a probe sonicator in nine 30 s intervals, with 30 s periods on ice at an amplitude setting of 40. 15 µL of sample was subjected to reverse cross-linking and electrophoresed on a 1% w/v agarose gel. Reverse cross-linking was achieved by adding 6 µL 5 M NaCl and 9 µL of water and incubating at 100 °C for 1 hour. The sample was subsequently incubated with 10 µg RNAse A and 18 µg Proteinase

K at RT for 15 minutes.

Pre-clearing of the prepared chromatin was performed by incubating 80 µL of Protein A agarose fast flow beads (Millipore, Cat No: 16156) , 1x complete protease inhibitor (Roche, Cat No: 04693132001), and Buffer (15 mM Tris-HCl pH 8.0, 165 mM NaCl, 20% Triton X-100, 1.2 mM EDTA) to a total volume of 600 µL at 20 rpm for 4 hours. Beads were subsequently separated by centrifugation and discarded.

8.5 µL of anti-CTCF (Millipore, Cat No 07-729) or anti-IgG (Abcam, Cat No: ab17890) or 3 µL of anti-Rad21 (Abcam, Cat No: ab992) was added to the chromatin and diluted to 600 µL with the aforementioned buffer. The solution was rotated overnight at 4 °C, transfered to Spin X columns (Costar, Cat No 8160), 60 µL of ProteinA beads were added and the columns were rotated for 2 hours at 4 °C.

Beads were washed in 800 µL of Buffer (20 mM Tris HCl pH 8.0, 150 mM NaCl, 0.1% SDS, 1% Triton X-100, 2 mM EDTA) by rotation at 4 °C for 15 minutes. The beads were washed again with the same buffer as above but with a total of 5 mM NaCl under the same conditions. A final wash was performed with buffer (10 mM Tris-HCl pH 8.0, 250 mM NaCl, 1% Igepal, 5% sodium deoxycholate, 1 mM EDTA). Beads were resuspended in 400 µL of water, 16 µL of 5 M NaCl were added and the solution was incubated overnight at 65 °C. DNA was purified by Phenol:Chloroform extraction.

2.8.2 Read Alignment and Identification of CTCF Enrichment Sites

This work was performed by Dr Reiner Schulz prior to this project and was also replicated by the author.

ChIP-seq read data were aligned to the mm9 built of the Mouse genome using Novoalign (version 2.07.11). Duplicates were removed using PicardTools `MarkDuplicates.jar` command. Peaks were identified at a permissive cutoff of FDR 0.5 using the USeq [Nix et al., 2008] package utilising the DEseq algorithm [Anders and Huber, 2010].

2.8.3 Filtering of CTCF and Cohesin Binding Sites

CTCF and Cohesin binding regions, with a Phred scaled [1] FDR exceeding 13 (95% calling accuracy), were selected independently using a custom UNIX shell script. The resulting peaks were expanded by 500 bp upstream and downstream and overlapping peaks were merged. All subsequent work was performed on this expanded and merged peak dataset.

[1] The Phred scale, mainly used for next-generation sequencing read scores is a log scale transformation in common use in Bioinformatics. The Phred score is defined as $Q = -10log_{10}P$ [Illumina, 2011]

2.8.4 Identification of Reads and SNPs within Regions of Interest

For performance reasons in subsequent processing the lists of SNPs and of reads were adjusted to include only the corresponding entries that overlapped or where fully within the CTCF or Cohesin sites, using the `interesectBed` command of Bedtools [Quinlan and Hall, 2010].

2.8.5 Mapping of Individual Reads to Alleles

Individual reads were assigned to alleles by means of a custom Perl script utilising the samtools Perl library [Li et al., 2009]. SNPs between the two parental strains were initially loaded to memory in one array per chromosome and sorted by genomic position. Aligned reads were read using the `samtools` Perl library and individually inspected for overlap with the polymorphisms by performing a binary search for the start position of the read in the SNP array and subsequently a linear search for its end. For all SNPs encountered in the interval, the matching parental allele was noted along with the sequencing score of the base pair, provision was made for single letter codes representing more than one base. The read was then assigned to a parental allele on the basis of the best base pair that overlapped a SNP. If no parental allele could be matched the read was recorded as unassigned. The above methodology was implemented in the script shown in Section C.1 of the Appendix.

2.8.6 Summarisation of Allele-specificity

Allele information from reads was first summarised per read pair. Every read pair was assigned according to the mapping of the read with the best SNP between the two strains present. If the read with the best SNP was unassigned, the read pair was assigned according to the mapping of the other read. Read pair mapping information was then adjusted for data obtained from CxB crosses so that the assignments were reflecting parental origin (maternal/paternal) as opposed to strain of origin. Counts of maternal and paternal reads were obtained, separately for each previously called CTCF and cohesin region. A probability value for the rejection of the null hypothesis (H_0), that the ratio of the reads observed a 1:1 ratio, was obtained by means of a two-sided binomial test. This was implemented in R [R Development Core Team, 2008].

2.8.7 Identification of Allele-specific Sites

Allele-specific sites were identified using a MySQL script. Data were loaded and sorted according to p-value. Allele-specificity was assessed by means of a Bonferroni corrected p-value cutoff, adjusted for the total number of regions examined for CTCF and Cohesin separately.

2.8.8 Identification of the CTCF Binding Motif

The MEME suite [Bailey et al., 2009] was used to confirm that the known CTCF binding motif was overrepresented in the identified CTCF binding sites. In particular the `meme-chip` tool was used to subsample the CTCF peaks and run the MEME motif identification tool on that sample. The analysis was also repeated using the best sub-window peaks identified by USeq – instead of the expanded and merged version of the peaks used throughout and described above. The repeated analysis reproduced the results with a higher lever of significance.

2.8.9 Assessment of Tissue-specificity of CTCF Binding Peaks

The tissue specificity of the CTCF peaks was assessed by comparison with CTCF peaks obtained in liver and ES cells. Data in the public domain were utilised [Schmidt et al., 2012] [Chen et al., 2008]. As the peak calling from liver and ES cell data was performed by different groups using dissimilar methodologies, it was considered necessary to standardise the peak size. To this end, peak size was optimised to identify the inflection point at which the majority of interactions are due to chance. In particular the peak size for all peaks was reduced to 1 bp and progressively increased. The number of peak overlaps for each step were counted and plotted. The optimal size of 1 kb was identified and used throughout. Peak overlaps were calculated using the `intersectBed` program from the BedTools suite with the '`-u`' switch. [Quinlan and Hall, 2010].

2.8.10 Identification of CTCF Peaks that do not Contain the Canonical Motif

Genomic sequence for the CTCF peaks was obtained using the Galaxy tool [Goecks et al., 2010] [Giardine et al., 2005]. The sequences were analysed with the FIMO tools (part of the MEME suite [Bailey et al., 2009]) to identify sequences that contained the canonical

CTCF motif as previously identified. These sequences were removed and their genomic coordinates obtained via custom UNIX shell scripts.

2.8.11 Identification of Putative Tissue-specific Binding Sites

The tissue-specific regions not containing the canonical motif, were prepared by applying the methodology described in Section 2.8.9 to the peaks not containing the CTCF canonical motif described in Section 2.8.10. These peaks were then processed with MEME to identify underlying motifs present exclusively in some tissues [Bailey et al., 2009].

2.8.12 Bisulphite Conversion and Cloning of *Magel2* promoter

The work described in this section was performed by Ms Siobhan Hughes.

Genomic DNA extracted from P21 BxC and CxB mouse brain tissue using the DNeasy Blood and Tissue Kit (QIAGEN, Cat No: 69504) was bisulphite converted using the EZ DNA Methylation-Direct Kit (Zymo Research, Cat No: D5020) according to the manufacturers' protocol. 2 µL of converted DNA was PCR amplified using the BSMagel F1 and R1 primers (Table A.1 of the Appendix) using Hot Star Taq DNA Polymerase (QIAGEN, Cat. No.: 203205) in a 25 µL reaction and incubated in a thermocycler using the protocol in Table 2.9.

Table 2.9: Colony PCR amplification

1	Initial activation and cell lysis	95 °C	15 minutes
2	Denaturation	95 °C	30 s
3	Annealing	53 °C	50 s
4	Extension	72 °C	30 s
5	Repeat from step 2, 44 times		
6	Final Extension	72 °C	5 minutes
7	Hold	4 °C	indefinitely

Following amplification, the PCR mix was electrophoresed on a 1.5% w/v LMP agarose gel on ice at 80 V for 90 minutes. DNA was recovered from the bands using the MinElute gel extraction Kit (QIAGEN, Cat No: 28604) according to the manufacturers' instructions. 20 ng of DNA were ligated to the pGEM-T Easy vector (Promega, Cat. No.: A1360) with overnight incubation. The DNA-vector construct was transformed into the completed *E. coli* cells (provided by Dr Sabrina Böhm) by incubation on ice for 30 minutes and heat shock at 42 °C for 45 seconds. The cells were returned to ice for a further of 2 minutes. 100 µL of S.O.C. medium (Invitrogen, Cat No: 46-0821) were added to

the cells and the cells were incubated at 37 °C for 1 hour with shaking. The cells were subsequently plated on Ampicillin (200 µL of 50 µg/mL in 100 mL), IPTG (8.4 µL of 1 M in 100 mL), X-Gal (100 µL of 40 mg/mL in 100 mL) plates and incubated overnight at 37 °C. Colonies were picked and incubated in 50 µL LB and ampicillin (0.05 µg/mL) at 37 °C for 2 hours with shaking, before grown overnight at room temperature.

2 µL of the growing culture was used directly in a PCR reaction with SP6 and T7 primers (Table A.1)(10 mM each) in 1.1X Reddy Mix (Thermo Scientific, Cat No: AB-06608/LD). The PCR program shown in Table 2.10 was executed

Table 2.10: Insert amplification PCR

1	Initial activation and cell lysis	96 °C	6 minutes
2	Denaturation	96 °C	60 s
3	Annealing	55 °C	90 s
4	Extension	72 °C	60 s
5	Repeat from step 2, 34 times		
6	Final Extension	72 °C	5 minutes
7	Hold	4 °C	indefinitely

The PCR products were electrophoresed on a 1% w/v agarose gel and the correct product size was confirmed. 2.5 µL of the PCR reaction mix was added to 10.5 µL of H_2O and 2 µL Illustra ExoStar (1 µL Illustra Alkaline Phosphatase and 1 µL Illustra Exonuclease) (GE Healthcare, US78220) and incubated at 37 °C for 15 minutes followed by 94 °C for 15 minutes.

Sanger Sequencing

Sequencing reactions were prepared on a 96 well plate by adding 2.5 µL of the previously treated DNA to 2 µL of 5X sequening Buffer, 0.4 µL of 10 µM Primer, 4.6 µL water and 0.5 µL Big Dye Terminator v3.1 (Invitrogen, Cat. No: 4337454) to each well. The sequencing reaction was performed by placing the plate in a PCR thermocycler and executing the program shown on Table 2.11.

DNA was precipitated by addition of 30 µL of 100% v/v ethanol and 1 µL of 3 M Sodium Acetate and incubation at 4 °C for 20 minutes. The plate was centrifuged at 3060 g at 4 °C for 20 minutes. The supernatant was removed and 30 µL of 70% v/v Ethanol were added before the plate was centrifuged again at 4 °C for 10 minutes. The supernatant was removed and the plate was allowed to dry at room temperature for 20 minutes. The DNA was resuspended in 10 µL of Hi-Di formamide (Applied Biosystems,

Table 2.11: Sanger sequencing reaction temperature cycle.

1	Initial activation	96 °C	1 minute
2	Denaturation	96 °C	30 s
3	Annealing	58 °C	15 s
4	Extension	62 °C	60 s
5	Repeat from step 2, 29 times		
6	Hold	4 °C	indefinitely

Cat No: 4311320) and 10 µL 1 mM EDTA was added to empty wells. The plate was incubated at 94 °C for 2 minutes, followed by a 5 minute incubation on ice before being sequenced on an ABI 3700xl sequencer.

Chapter 3

Epigenetic and Transcriptional Regulators of Endocardial Cells

The heart is made of three distinct tissue layers, the endocardium, the myocardium and the epicardium. The inner endocardial layer is an endothelial population of cells lining the ventricles and atria of the heart that is known to have an important role in the development of the heart. Despite its importance, the endocardium is relatively understudied compared to the myocardial population.

Interest in the endocardium stems from the fact that it has an essential role in heart formation and in particular valve formation, trabeculation of the ventricles and transdifferentiation of cardiac muscle to Purkinje conduction fibres via reciprocal interactions with the myocardium. Given that the aetiology of congenital malformations of the heart, especially that of valvular defects, is poorly understood and that the pathways that lead to these defects are likely to involve endocardial cells, this is a cell population of great interest.

Although the origin of the endocardial population is known to be in the lateral plate mesoderm, its exact relation to other heart cell populations is not clear (Section 1.3.3). Studies have shown that it separates from the myocardial population before formation of the heart tube but its exact origin remains elusive. The existence of a cardiogenic progenitor cell that is able to give rise to all heart populations, including endocardium, is well described [Moretti et al., 2006] [Kattman et al., 2006], but the extent to which endocardial cells in the embryo originate from this precursor has not been defined. Furthermore, conflicting data exist suggesting that the endocardial cells originate from a vascular endothelial precursor [Milgrom-Hoffman et al., 2011]. Other evidence suggest

that the endocardium can give rise to coronary arteries, further complicating our understanding of these cells [Wu et al., 2012].

Between E8.5 and E9.5 of development, endothelial cells can be identified by the expression of the NFATc1 transcription factor [Misfeldt et al., 2009]. The presence of a distinct marker at a specific developmental stage suggests that the endocardial cells are to a considerable extent a functionally uniform population at that time point. Furthermore, the double knockout of two related tyrosine kinase receptors ($Tie1^{-/-}$ $Tie2^{-/-}$) expressed in all vasculature during late development and in the adult displays cardiac specific defects but no early vascular defects, suggesting a specific identity for the endocardium [Harris and Black, 2010].

Despite its ubiquitous presence in the endocardium and its function as a transcription factor, NFATc1 is unlikely to be a master regulator of endocardial identity. The NFATc1 double knockout mice do not show absence of the endocardium, although they show aberrant valve formation [Ranger et al., 1998]. NFATc1 expression does not precede the specification of the endocardium and is furthermore expressed in other tissues during development, such as cartilage during limb development [Misfeldt et al., 2009] and in the adult it has a role in the the activation of T-cells.

Epigenetic and transcriptomic characterisation of the endocardium was undertaken in the course of this project. The embryoid body cardiac differentiation model was used to isolate endocardial and endothelial cells for analysis in sufficient quantities. The embryoid body differentiation model of endocardial cells has been shown to recapitulate endogenous temporal and spatial expression patterns [Misfeldt et al., 2009] (Section 1.3.5).

Cells bearing the $NFATc1^{+}/CD31^{+}$ endocardial signature were isolated from embryoid bodies with flow cytometry and compared to endothelial cells ($NFATc1^{-}/CD31^{+}$) from the same source. Endothelial cells were used as a background population for comparison as they represent a cell type of identical morphology and similar function, that does not however possess the unique properties of the endocardium, such as the ability to undergo EMT or induce trabeculation. Furthermore, endothelial cells can be obtained in quantities equal to or greater than endocardial cells from mouse embryoid bodies.

Our initial hypothesis was that epigenetic processes have a major role in endocardial cell identity. This hypothesis was based on unpublished differential expression microarray analyses performed prior to and outside the scope of this project that suggested no transcriptional differences between endocardium and endothelium (Prof. Scott Baldwin,

Personal Communication). Furthermore, at the time point examined (E9.5 equivalent), the endocardial and endothelial cells are morphologically identical, but manifest differential responses to exogenous stimuli later in development, suggesting the existence of hidden variables that may be of epigenetic identity.

3.1 Cell Isolation

3.1.1 Isolation of Cells from Embryoid Body Cultures

Isolation of endocardial and endothelial cells from embryoid bodies was performed by members of the Baldwin Lab, TN, USA and it is presented here for reference purposes. Briefly, embryonic stem cells that were transfected with the NFATc1-mCherry BAC, derived from the NFATc1-nuc-LacZ BAC transgene reported by Misfeldt and colleagues [Misfeldt et al., 2009], were propagated on primary embryonic fibroblasts and differentiated in embryoid body cultures for two days. After culture on a rotary orbital shaker for 48 hours, the cells were plated on gelatine. Plated cells were allowed to differentiate for at least 7 days before being harvested. Detailed experimental procedures can be found in Section 2.1.

Endocardial and endothelial cells were disassociated and isolated by flow cytometry on the basis of the presence of the mCherry (expressed under the control of NFATc1 promoter and enhancer) and CD31 markers. Endocardial cells were isolated as the $NFATc1^{+}/CD31^{+}$ population and endothelial cells as the $NFATc1^{-}/CD31^{+}$ population. Representative FACS isolation plots for individual markers and doubly labelled cells can be seen in Figure 3.1.

3.2 Genome-Wide Methylation Analysis

3.2.1 Establishment of the WGBS protocol

A protocol for performing whole genome bisulphite sequencing was established using the published whole genome reduced representation bisulphite sequencing (RRBS) as a starting point [Meissner et al., 2005]. The RRBS protocol generates bisulphite converted sequencing libraries enriched for CGIs by digestion using a methylation insensitive enzyme, followed by size selection for a fraction with high representation of CGIs.

The enzymatic digestion step was substituted with acoustic shearing, to ensure un-

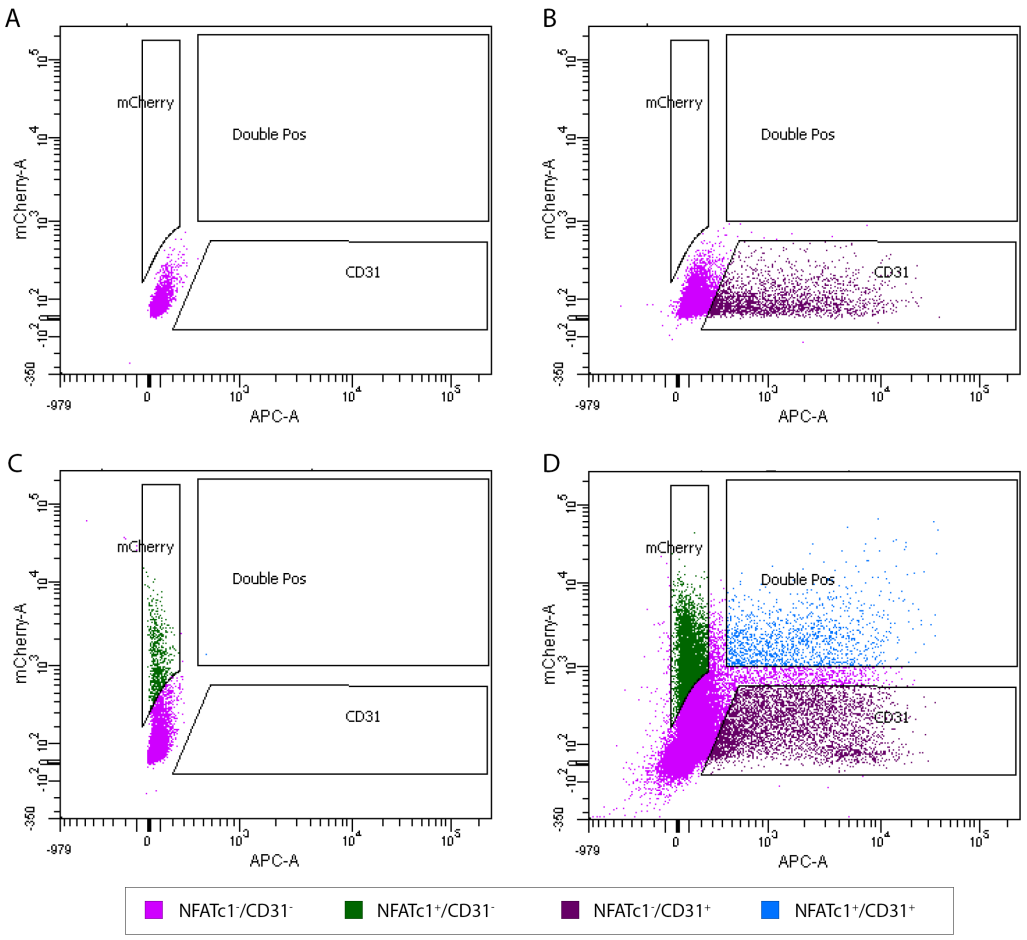


Figure 3.1: Representative FAC sorting plots provided by the Baldwin Laboratory. A. Sorting of unlabelled cells B. Sorting of CD31 labelled cells C. Sorting of mCherry labelled cells D. Sorting of double labelled cells. Colour is used to signify the sorting designation of each cell.

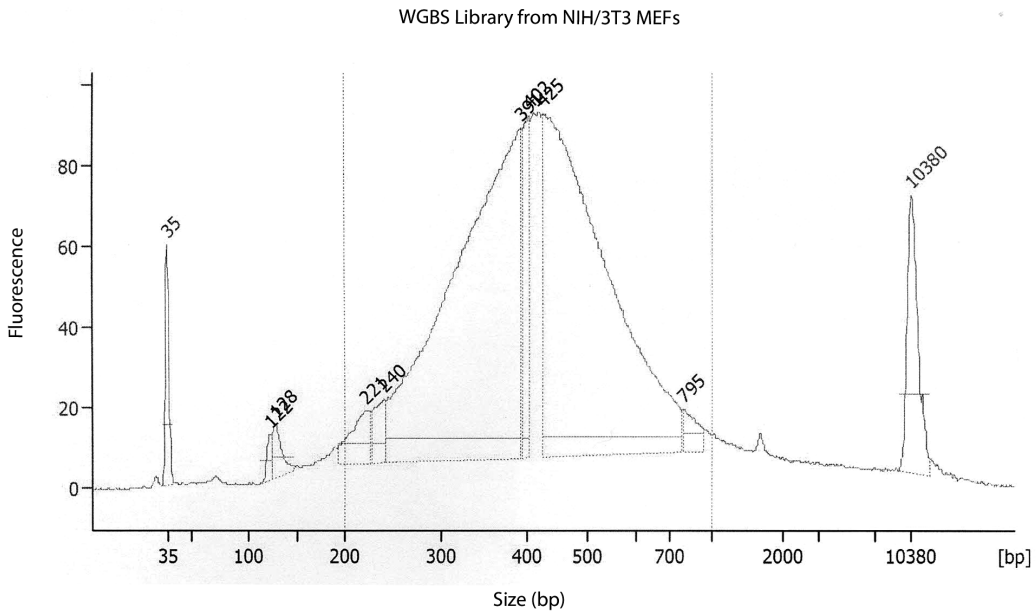


Figure 3.2: Bioanalyser trace of WGBS next-generation sequencing library prepared with NIH/3T3 MEF DNA for protocol validation. The trace shows a distinct peak at 400 bps. The peaks at 122 bp and 10,380 bp are lower and upper marker peaks respectively used for size estimation.

biased and uniform fragmentation, and next-generation sequencing library preparation was performed using a commercially available library preparation kit. A flowchart of the protocol can be found in Figure 3.3. After sonication the DNA was end-repaired, A-tailed and methylated adaptors were ligated. The library was subsequently bisulphite converted using a commercially available kit and amplified in two steps. Note that only half the bisulphite converted DNA is used for amplification, with the other half stored for later usage.

The WGBS protocol was first attempted on NIH/3T3 mouse embryonic fibroblasts (MEFs) to a favourable outcome (Figure 3.2) before proceeding to use it on DNA from endocardial and endothelial cells.

3.2.2 Experimental Design and Library Preparation

At the commencement of this project, two batches of DNA from flow sorted embryoid bodies for each of endocardial and endothelial cells were available. The samples were quantified after a 1:10 dilution of 1 µL of each sample using the quBit instrument (Table 3.1) and the total amount of DNA in each sample was calculated (Table 3.2). The total amount of DNA was judged not to be sufficient for WGBS on the basis of previous

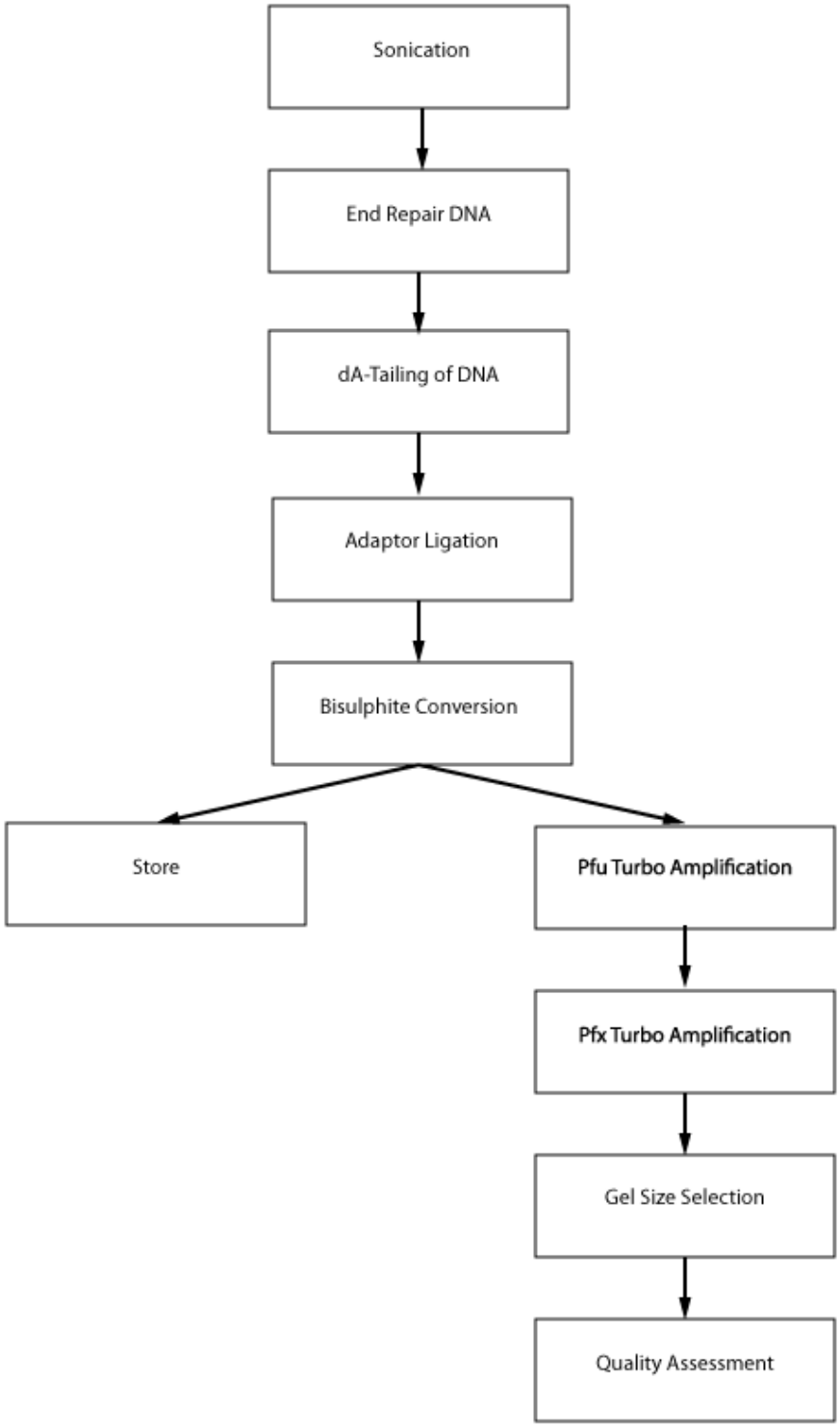


Figure 3.3: Flowchart of WGBS library preparation protocol. The DNA is sheared by sonication and a next generation sequencing library is prepared by end-repairing, dA-tailing and adaptor ligating. Custom made methylated adaptors are used during adaptor ligation. The library is bisulphite converted and half the sample is used for subsequent processing, the remaining sample is stored. Amplification of the library fragments is performed by two consecutive rounds of PCR amplification. The first round is performed using Pfu DNA polymerase, an enzyme that can read through U. The second round is performed with Pfx DNA polymerase. DNA of the appropriate size is selected by gel electrophoresis and quality is assessed. DNA purification is performed between the shown steps.

attempts with NIH/3T3 fibroblasts. The samples were therefore pooled and quantified again. Table 3.3 shows the mass, volume and expected and actual concentrations of samples after pooling. The close correspondence of the actual and expected concentrations confirms that pooling has been performed correctly.

Four libraries were prepared from the first replicate samples, three of which were sequenced on seven Illumina ®Hiseq 2000 lanes (Table 3.4). Libraries EC0 and ET1 were prepared in a single batch. Pooled DNA was sonicated and its size distribution assessed using the Agilent Bioanalyser (Figure 3.4). The Bioanalyser trace for ET1 displayed a sharp peak at a molecular weight of 374 bp, but the broad peak of the DNA was otherwise unaffected. The sharp peak was attributed to an air bubble in the capillaries and was consistent with similar observations by other users of the instrument.

Library EC0, in comparison to library ET1, was judged to be unsuccessful given that it was nearly undetectable on the Tapestation analysis (Figure 3.5). Subsequent re-analysis of the quantification data revealed that this library could have been sequenced (Table 3.5 compare EC0 with ET2), although it was indeed suboptimal. Library EC1 was prepared to replace library EC0 from bisulphite converted DNA remaining from the preparation of library EC0 (only the post-bisulphite amplification steps were repeated). During preparation of libraries EC0 and ET1 it was noticed that the size selection resulted in fragments suboptimal for 200 bp paired-end sequencing and the size-selection step was adjusted appropriately for library EC1 (Figure 3.5). Library ET2 was prepared from bisulphite converted DNA from the ET1 library preparation to optimise the insert size. These libraries (ET1, EC1 and ET2) were all prepared using DNA from the sample pool of embyoid cells and were therefore treated as technical replicates in subsequent analysis.

Two further libraries (one from endocardial and one from endothelial DNA) were prepared from independent pools of cells that became available later in the course of the project constituting a biological replicate. These libraries are identified as EC2 and ET3. Preparation of these libraries was performed by Mr Samuele Maria Amante, using the protocol developed by the author. The molar concentration of these libraries was calculated using a qPCR assay and the fragment size information shown in Figure 3.5.

The molar concentration of all libraries was calculated (Table 3.6). Libraries were diluted to 10 nM before being sequenced.

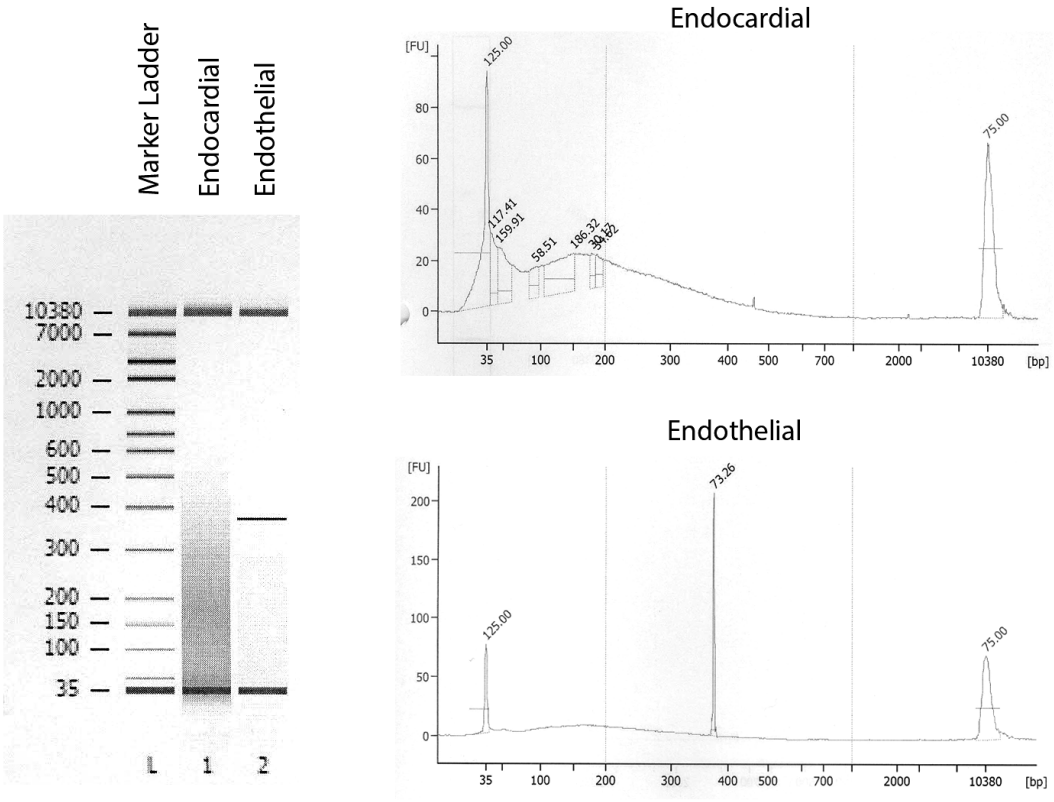


Figure 3.4: Bioanalyser gel (A) and trace representation (B,C) of fragment size analysis for endocardial and endothelial DNA after sonication and prior to library preparation. Both libraries show peaks of comparable size at around 200 bp, suitable for library preparation (note different scale). The endothelial library shows a sharp spurious peak slightly smaller than 400 bp, see text for more details.

Table 3.1: Quantification of diluted DNA samples used for replicate 1 of the WGBS and calculation of original concentration. Endocardial sample 2 was more concentrated than other samples. This discrepancy was consistent with the sample quantification by our collaborators after the DNA isolation and prior to shipping (data not shown).

		Concentration (µg/mL)		
		Read 1	Read 2	Mean
Endocardial 1	Diluted	0.256	0.245	
	Original	2.56	2.45	2.51
Endocardial 2	Diluted	1.67	1.64	
	Original	16.7	16.4	16.6
Endothelial 1	Diluted	0.342	0.367	
	Original	3.42	3.67	3.55
Endothelial 2	Diluted	0.201	0.202	
	Original	2.01	2.02	2.02

Table 3.2: Calculation of total amount of DNA present in samples of replicate 1 prior to pooling. Some samples contained less than 50 ng of DNA that was considered the minimum required for library preparation.

	Concentration (ng/µL)	Volume (µL)	Total DNA (ng)
Endocardial 1	2.51	19	47.7
Endocardial 2	16.6	19	315.4
Endothelial 1	3.55	19	67.5
Endothelial 2	2.02	19	38.4

Table 3.3: Total DNA calculation of DNA samples used for WGBS after pooling.

	m_T (ng)	V_T (µL)	$c_{expected}$ (ng/mL)	c_{actual} (ng/mL)
Endocardial	363.1	38	9.56	7.75
Endothelial	105.9	38	2.79	3.11

Table 3.4: Summary, number of lanes sequenced and status of prepared libraries for WGBS of endocardial and endothelial cells.

Library	Cell type	Status	Lanes Count	Biological Rep.	Notes
EC0	Endocardial	FAIL	0	1	dilute; small insert
ET1	Endothelial	OK	1	1	small insert size
EC1	Endocardial	OK	3	1	
ET2	Endothelial	OK	3	1	
EC2	Endocardial	OK	0.5	2	
ET3	Endothelial	OK	0.5	2	

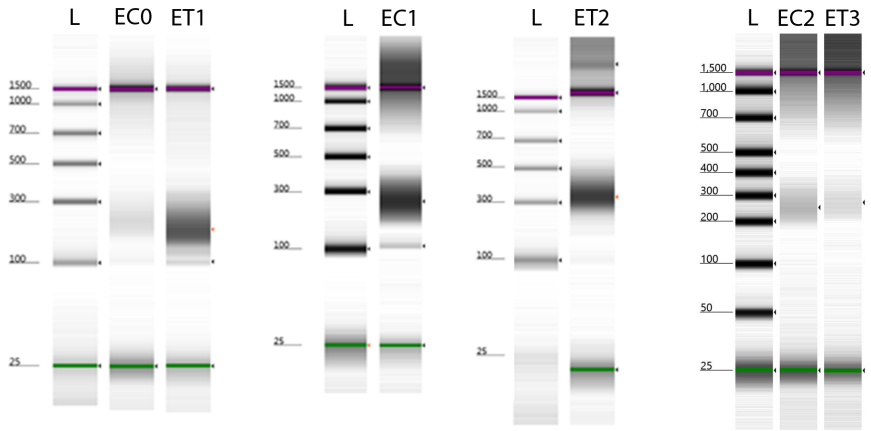


Figure 3.5: Estimation of the size of WGBS libraries using the Agilent Tapestation. The size distribution is displayed as a gel. All libraries are of the appropriate size. Library EC0 is clearly more dilute than ET1. Library ET1 was smaller than the ideal size and was later repeated as ET2, see text for details.

Table 3.5: DNA quantification of 1:200 or 1:2000 dilution of prepared WGBS libraries using the quBit. All libraries contained enough DNA for sequencing.

	Dilutions	Read 1	Read 2	Mean
EC0	1:200	8.48 ng/mL	9.87 ng/mL	
	original	1.70 µg/mL	1.97 µg/mL	1.84 µg/mL
ET1	1:200	32.9 ng/mL	33.9 ng/mL	
	original	6.58 µg/mL	6.78 µg/mL	6.68 µg/mL
EC1	1:2000	3.95 ng/mL	4.01 ng/mL	
	original	7.89 µg/mL	8.02 µg/mL	7.96 µg/mL
ET2	1:2000	1.04 ng/mL	0.98 ng/mL	
	original	2.08 µg/mL	2.00 µg/mL	2.04 µg/mL

Table 3.6: Calculation of molar concentration of WGBS libraries. The molar concentration for EC0 was not calculated, as a reliable insert size estimate was not available. The concentrations of libraries EC2 and ET3 was determined directly via qPCR.

Library	Concentration (µg/mL)	Insert Size (bp)	Concentration (nM)
EC0	1.84	N/A	N/A
ET1	6.68	183	55.31
EC1	7.96	250	48.24
ET2	2.04	326	9.48
EC2	N/A	275	158.38
ET3	N/A	283	102.57

3.2.3 Sequencing Results

Cluster generation and sequencing was performed by the BRC genomic facility as outlined in Section 2.3.2.

One lane of the ET1 library was initially sequenced before proceeding to sequence three lanes of each EC1 and ET2 as shown in Table 3.4. In addition, one multiplex lane of a 1:1 mix of the EC2 and ET3 libraries was sequenced. Sequenced lanes are identified by the library ID and a serial number indicating the number of times the originating library has been sequenced. For example, ET2-3, indicates the third lane sequenced from the second library from endothelial cells. Total cluster and read counts obtained are shown in Table 3.7. The depth of sequencing was based on that employed by other WGBS experiments and personal communication with experts in these methodologies (Dr. Miguel Branco, personal communication).

An instrument failure occurred during the sequencing of lanes EC1-2 and ET2-2 leading to termination of sequencing before completion of all cycles for the reverse read, this led to 100 bp forward reads but only 60 bp reverse reads for those lanes. Libraries of the second biological replicate were multiplexed and sequenced on a single lane.

Table 3.7: Read pair count for whole genome bisulphite sequencing of endocardial and endothelial cells prior to quality control.

Lane Identifier	Endocardial Read Pair Count	Endothelial Read Pair Count
EC1-1	251,478,868	0
EC1-2	259,260,116	0
EC1-3	237,855,069	0
ET1-1	0	174,340,409
ET2-1	0	279,379,754
ET2-2	0	291,248,642
ET2-3	0	280,003,668
EC2-1	98,063,821	0
ET3-1	0	91,772,411
Total	846,657,874	1,116,744,884

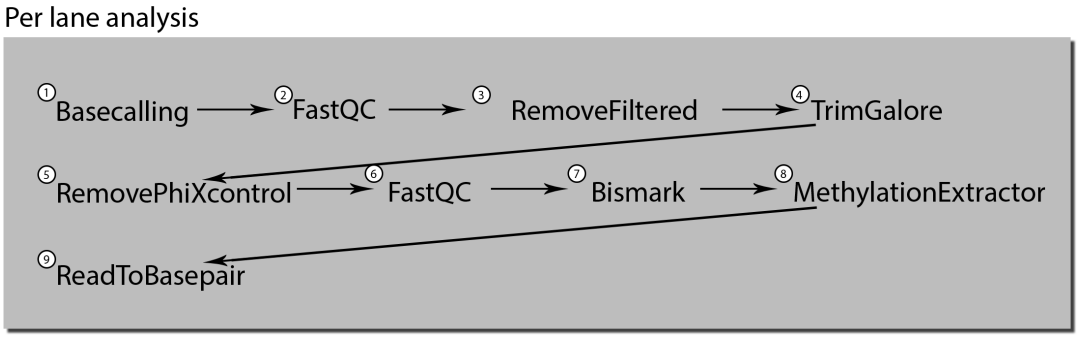


Figure 3.6: Simplified outline of the per-lane WGBS pipeline developed in the context of this project. Some quality evaluation steps have been omitted for clarity. See main text page 101 for details.

3.2.4 Bioinformatic Processing

The bioinformatic processing pipeline developed and utilised is shown diagrammatically in Figure 3.6. The numbers in the following text refer to step numbers in that Figure. Individual lanes were treated as technical replicates and analysis was initially performed on a per-lane basis and subsequently data were summarised on a per-feature basis. Details of the processing can be found in Section 2.3.3.

Basecalling was performed with Casava (1) and read quality was initially assessed with FastQC (2) (Figure 3.7). The initial quality assessment revealed a large range of per base quality values with several lanes including reads of low quality. This was attributed to overloading of the sequencer lane.

The basepair composition was found to be consistent with bisulphite conversion, with Cs underrepresented and Ts overrepresented. This trend however was not consistent throughout the read length, with the basepair composition approximating 25% towards the end of the reads. This suggested read-through into the unmethylated adaptor, which was consistent with adaptor sequence overrepresentation identified by FastQC.

Reads marked as filtered by the sequencer instrument were removed (3). The sequencer instrument marks reads from removal on the basis of proximity and overlap of clusters as well as lack of sequence complexity. Given the overloading of some of the flowcells, this was therefore judged to be necessary, despite the considerable loss of clusters at this step (see Figure 3.10). Quality was assessed again (not shown in Figure 3.6).

Adaptor trimming was performed with `trim_galore` (4). Adaptor trimming was deemed to be essential given the increase in Cs towards the ends of the reads as described above. A total of 14.4% of the total number of bases were trimmed from 62.7%

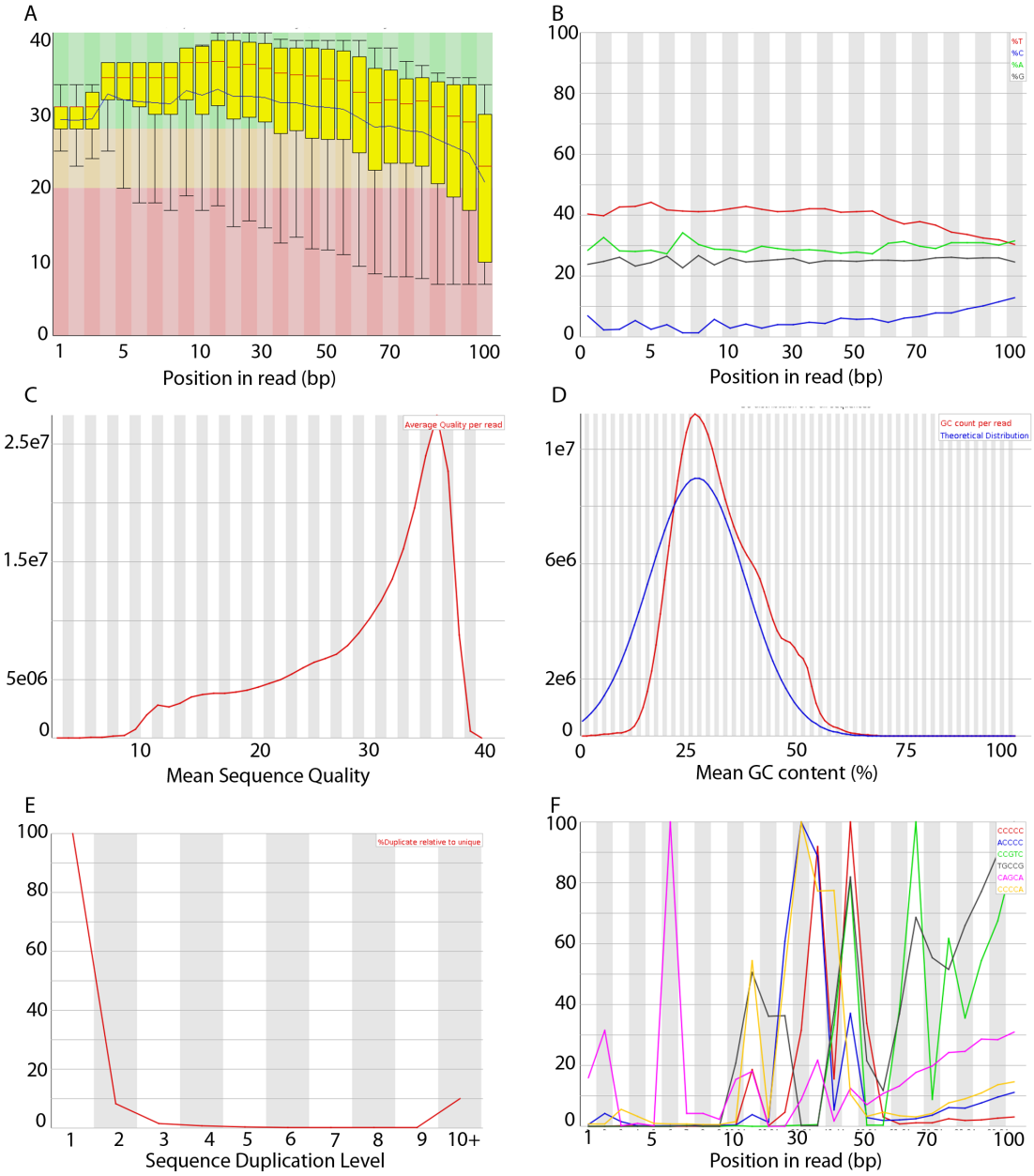


Figure 3.7: Quality plots of raw forward reads, prior to quality control, for lane EC1-1. These plots are broadly representative of other lanes, although some lanes did show considerably lower quality, see text. (A) Phred scaled quality score box plot as a function of read position suggests good overall read quality and displays the expected quality drop towards the 3' end. (B) Base incorporation percentage as a function of read position, suggests bisulphite conversion was successful (overall percent of Cs lower than expected and overall % of Ts higher than expected). The cycle-to-cycle incorporation variability suggests either an instrument malfunction or overrepresentation of specific sequences in the library. Increasing C and decreasing T percentage suggests reading into the methylated (and unconverted) adaptor. (C) Average read quality per read is high, but displays a long tail of low quality reads. (D) GC% distribution does not match the expected distribution as expected from WGBS. (E) Duplication level of reads, shows a moderate to high duplication level, which may be attributable to overrepresented sequences. (F) K-mer content as a function of read position, shows overrepresentation of several k-mers in a position specific manner.

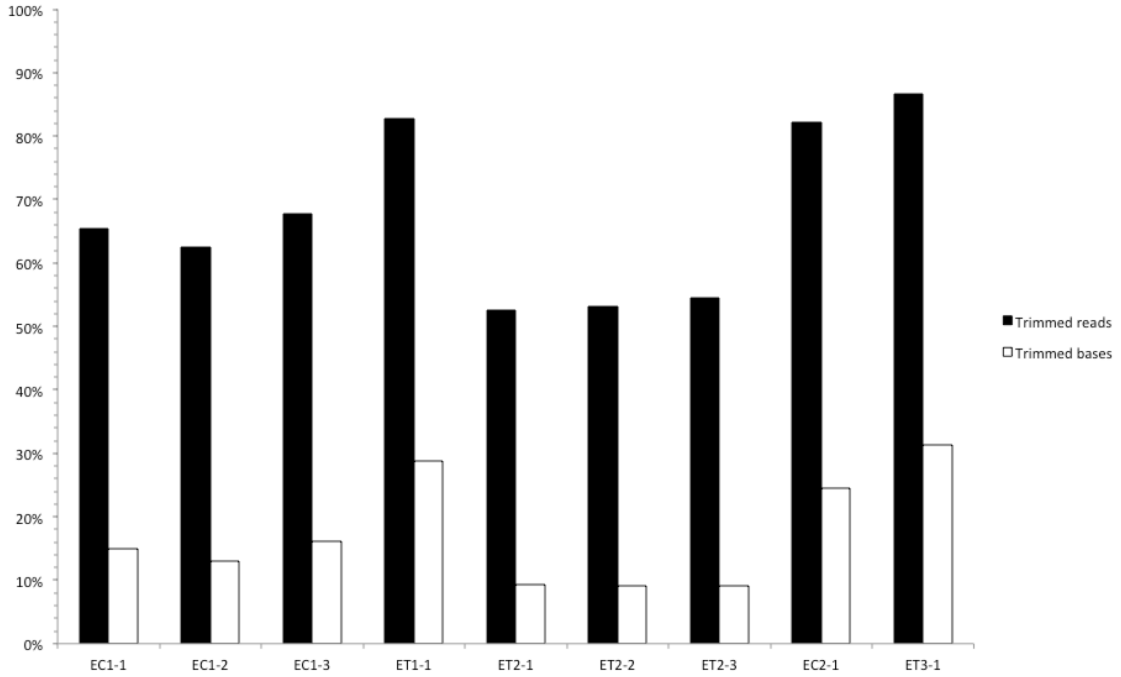


Figure 3.8: Percent of total trimmed reads and total trimmed bases. A considerable portion of all the reads was trimmed to some extent, while the overall percent of bases trimmed was lower than 30% for all libraries.

of the total reads (see Figure 3.8).

Trimmed reads were aligned to the phiX reference genome to remove phiX DNA added for quality control purposes after library construction and prior to sequencing by the sequencing facility (5), reads that successfully aligned to the phiX genome were removed. The contribution of the phiX reads to the total read count was small (see Figure 3.10). This step was not performed on the multiplexed samples as phiX reads were not barcoded and were discarded during the demultiplexing step.

Quality was evaluated again with FastQC, see Figure 3.9 (6). By comparing Figures 3.8 and 3.9 it is evident that the quality control steps have resulted in improvement of the data quality. Phred scaled per-base quality scores are consistently above 29 (>99.9% accuracy) (panels A), sequence composition is stable throughout the read (panels B), reads with low mean quality scores are absent (panels C), GC distribution resembles the expected distribution with a second mode at 61% methylation (panels D), that is interpreted as the methylated (unconverted) component of the genome, duplication estimates are below 20% and the majority of duplication is of low copy number (panels E) and k-mer content is considerably more stable throughout the reads (panels F).

The reads were bisulphite converted *in silico* and aligned to the *in silico* bisulphite converted genome, using Bismark [Krueger and Andrews, 2011] (Section 2.3.3).

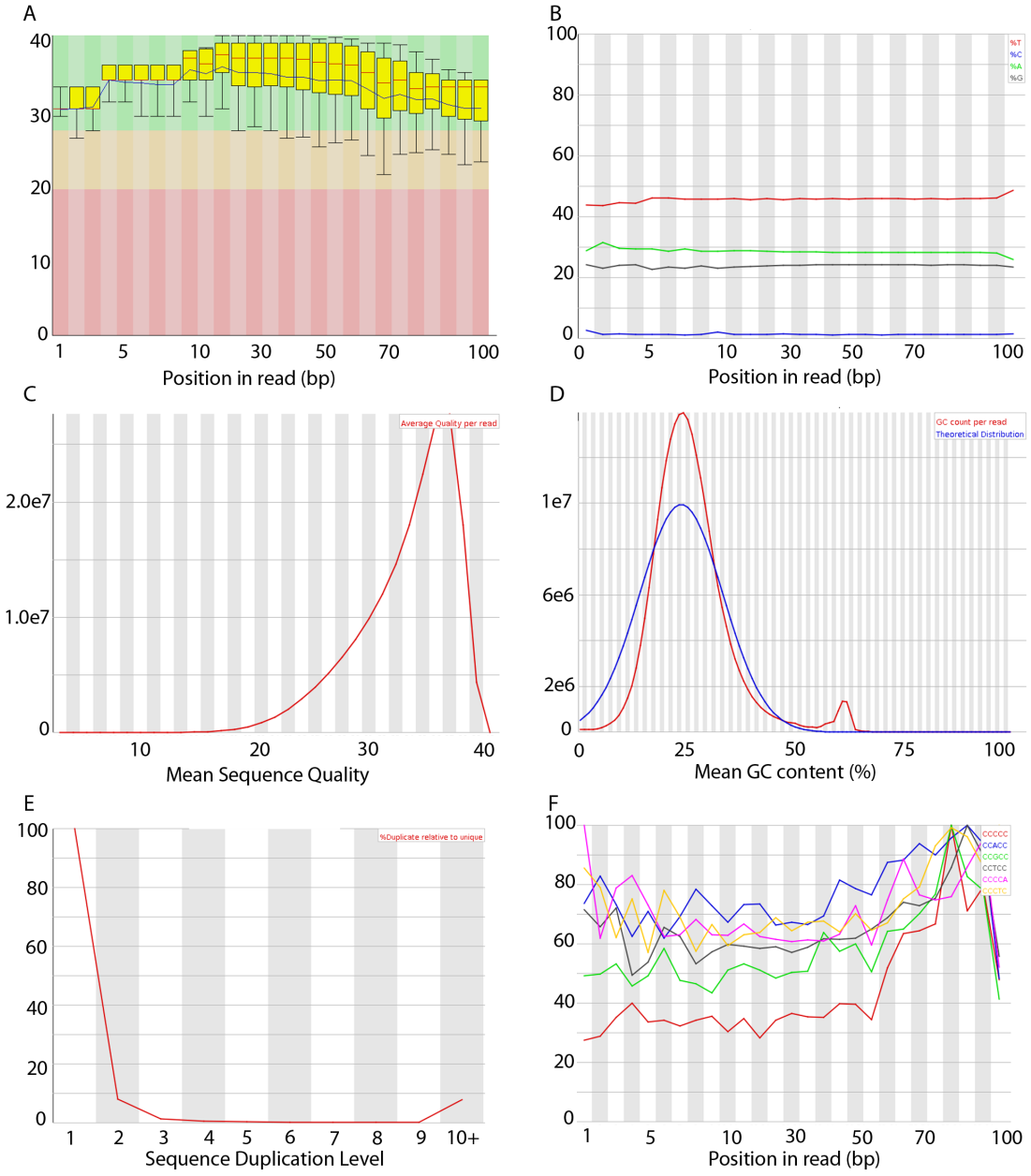


Figure 3.9: Quality plots of raw forward reads, following quality control, for lane EC1-1, compare with Figure 3.7 on page 102. These plots are representative of other lanes. (A) Phred scaled quality score box plot as a function of read position shows consistently high and improved quality scores. (B) Base incorporation percentage as a function of read position does not show significant cycle-to-cycle variation and no overall trend towards the 3' end of the read is evident supporting successful removal of the majority of the adaptor sequences. (C) Average quality per read distribution plot shows improvement compared to pre-quality control. (D) GC% distribution shows a clear bimodal distribution, potentially corresponding to methylated and unmethylated genomic regions. (E) Duplication rate is considerably lower and furthermore shows reduction of reads with more than 10 multiple copies. (F) K-mer content as function of read position is now consistent throughout reads with no abrupt changes suggesting that overrepresented sequences are of low abundance and no significant positional fragmentation bias remains.

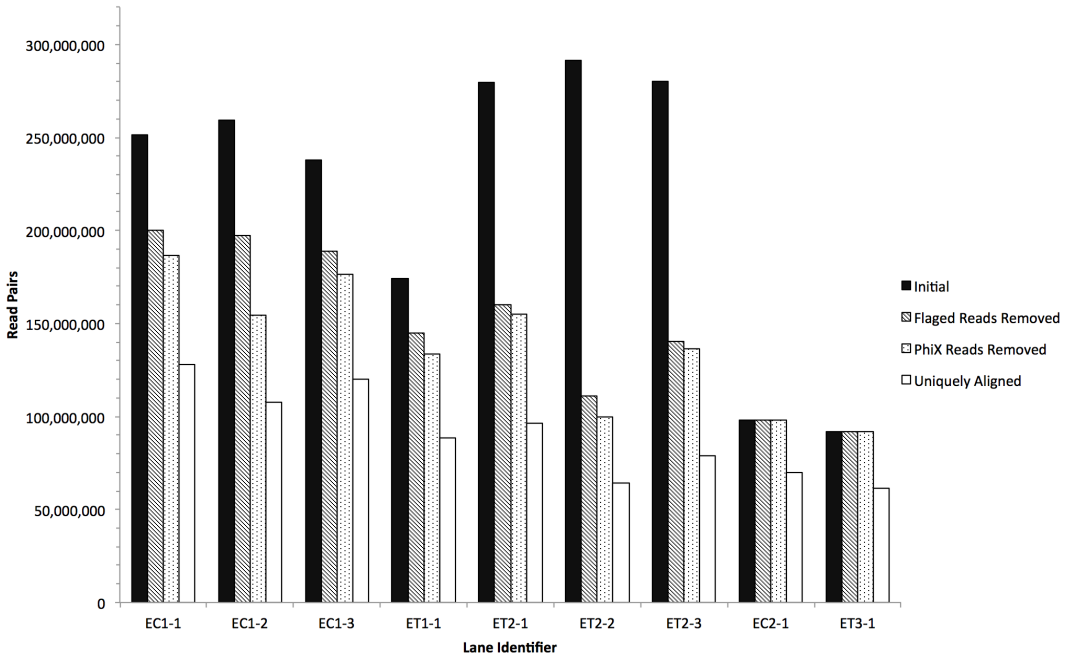


Figure 3.10: Barchart of read counts for WGBS for several steps of the analysis. The majority of lost reads were attributable to removal of reads filtered by the sequencer, this was due to the overloading of the flowcell that resulted in overlapping clusters.

De-duplication was attempted with two different approaches but was not performed in the final version of the processing pipeline. This was because de-duplication with Picard tools resulted in files that did not contain the appropriate flags from `Bismark` required for methylation extraction. Furthermore, de-duplication using a custom experimental script packaged with `Bismark` was not successful due to the very high memory requirements of the script that rendered it impractical to complete its execution. Given that duplication rates following quality control were estimated not to be excessively high (Figure 3.9) and the aforementioned technical difficulties, this step was omitted and no significant adverse effects are expected because duplication rates were low.

The methylation information from all reads was obtained (8) and summarised on a per base-pair position (9) using a custom script. Summarisation per-base position prior to summarisation per feature was necessary to ensure that unequal coverage of Cs within each feature does not result in preferential weighting of highly covered C positions.

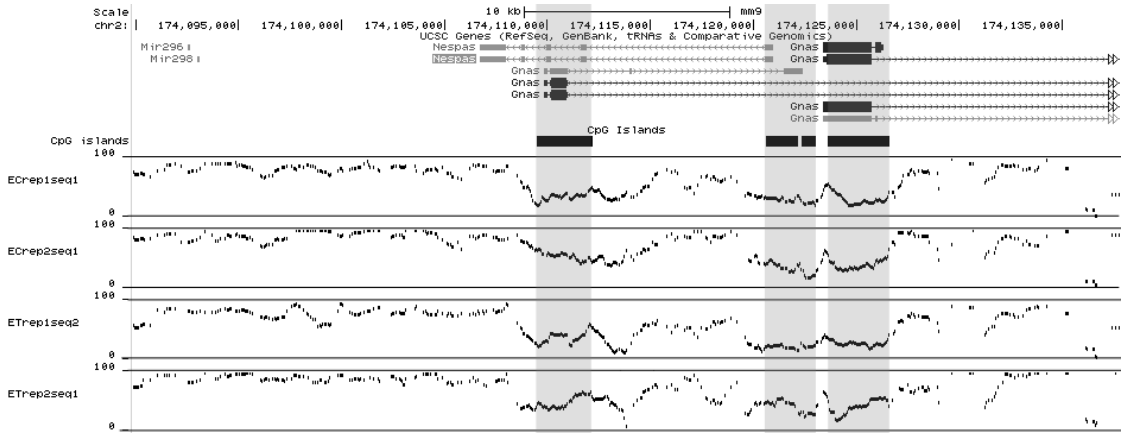


Figure 3.11: Overview of methylation at the known imprinted *Gnas* locus displaying hemimethylated CGIs. Hemimethylation of imprinted loci suggests that the methylation profile of the cells examined is not perturbed and is relevant to the *in vivo* context.

3.2.5 Visual Inspection

The data were visually inspected on the UCSC Genome Browser at the read and methylation level. The read data were not found to exhibit read patterns indicative of significant duplication, displayed the anticipated mismatches to the genomic sequence at the majority of C residues and coverage appeared to be uniform. The methylation pattern was consistent with a hypermethylated genome and hypomethylated CpG islands.

3.2.6 Methylation at Imprinted Loci

Methylation at imprinted loci, where methylation levels are expected to approximate 50% was examined at each locus. The majority of the loci exhibited methylation patterns consistent with the methylation of only one of the two alleles. An example of this can be seen in the *Gnas* locus (Figure 3.11).

The methylation of all the imprinted loci was quantified and is presented in Figure 3.12. The majority of the loci are in a hemimethylated state. Only three loci exhibited a methylation state incompatible with hemimethylation (*Slc38a4*, *Rasgfr1* and *Peg13*).

The reason for loss of the expected methylation pattern at these loci could either be due to a previously unknown tissue specific loss of allele-specific methylation in endothelial cells at these sites or alternatively due to the loss of the normal methylation pattern at these loci in the particular ES cell line used in this analysis.

The presence of the expected methylation patterns in the majority of the examined imprinted loci and expected pattern of CpG methylation throughout the genome serves as an internal control and demonstrates that the epigenetic profile of these cells is not

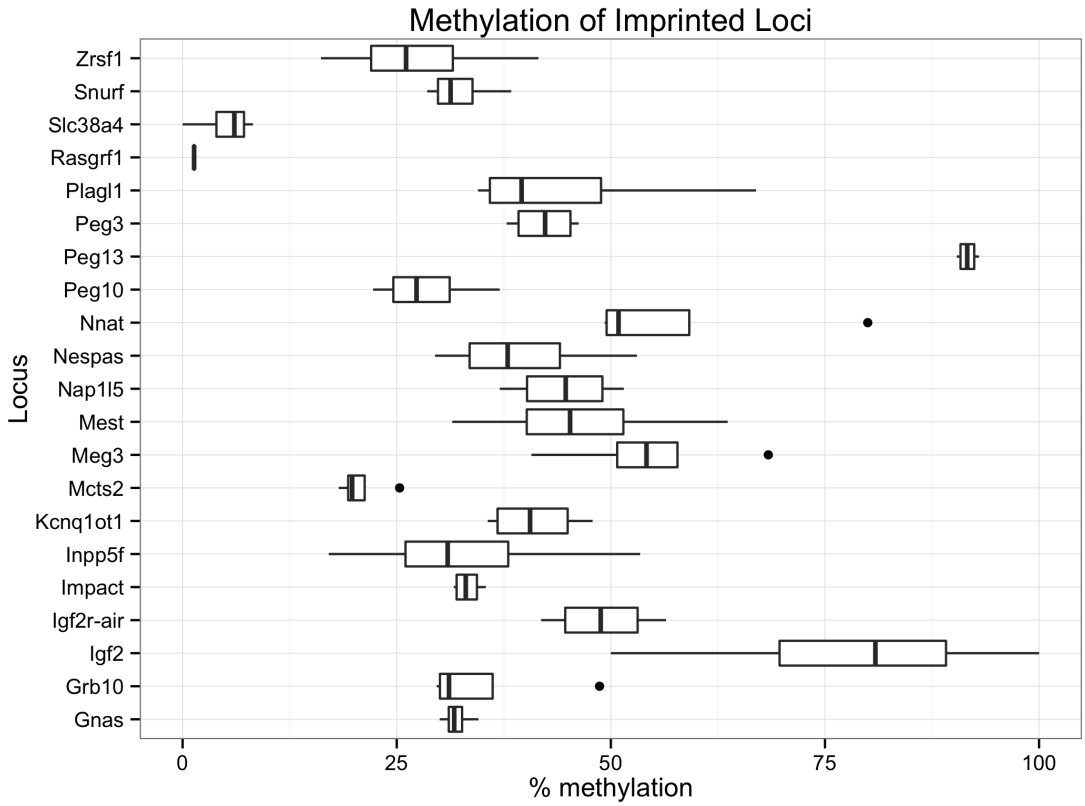


Figure 3.12: Summary of methylation of imprinted loci. The majority (16/21) of the loci are in an hemimethylated state. Only three of the remaining loci exhibit methylation suggestive of a completely methylated or unmethylated state (*Slc38a4*, *Rasgfr1* and *Peg13*). These data support the notion that the epigenetic state of the cell cultures is largely unperturbed.

significantly altered as a result of the growth conditions. This supports the notion that the cells are appropriate for the analysis.

In the course of the above analysis, a novel region exhibiting hemimethylation was identified at the vicinity of *Inpp5f* locus in both tissues. Although this study did not generate allele-specific data, the proximity of this hemi-methylated region to a known DMR suggests that it may be methylated in an allele-specific manner. The biological significance of this finding was not evaluated further as it was outside the scope of the present investigation.

3.2.7 Differential Methylation of CGIs

Data from the WGBS experiment were processed as described in Section 2.3.3. Briefly, reads were subjected to quality control, aligned to the *in silico* bisulphite converted genome and methylation information was extracted and summarised to the CpG level.

The coverage distribution of CpGs in each sample was examined (Figure 3.13 A and B) and a minimum coverage cutoff of 3 reads was set. Furthermore CpGs with coverage in excess of 100 were excluded in order to exclude regions with high levels of duplication. This was considered necessary given than per read level duplicate removal had not been performed, due to technical constraints.

The data were summarised to the CGI level. A script was developed (initially with MySQL and later with bedmap [Neph et al., 2012]) to perform the summarisation.

The dataset of known CGIs was prepared by producing the union of the datasets of CGIs obtained from UCSC genome browser CGI track [Gardiner-Garden and Frommer, 1987] and CGIs identified by Bird and colleagues via CAP-seq [Illingworth et al., 2010]. Overlapping regions between the two datasets were merged. These two datasets were used as Illingworth and colleagues demonstrated that a portion of the functional CpG islands in the mouse have not been computationally detected and can be detected by means of CAP-seq. This dataset was not however used in isolation to avoid exclusion of well described CGIs that were not identified in this particular assay.

The distribution of the number of informative CpGs per CGI were examined (Figure 3.13, C). Only CGIs with at least 10 informative CpGs were considered for further analysis. Filtering of CGIs was performed so as to remove CGIs where differential methylation cannot be reliably detected and reduce the extent to which multiple testing correction required.

The methylation of individual CGIs between the two conditions was examined by means of a scatter plot (Figure 3.14). Of particular interest was the overall deviation from the 45 degree line that was observed. This trend suggests overall hypermethylation in the endocardium across many CGIs. This hypermethylation is consistent with a model of differentiation whereby endocardial cells are a more differentiated subtype of the endothelium, gaining methylation as they differentiate. Alternatively, this may signify a considerable contribution of undifferentiated hypomethylated cells to the CD31+/NFATc1- population. The later possibility is partly supported by the upregulation of some pluripotency factors in the endothelial population such as Nanog observed in the transcriptomic analysis of this population (see page 142).

The possibility existed however, that this overall difference in methylation between the two conditions was the result of the differential conversion efficiency between the two samples, that manifests as differential methylation. The conversion efficiency was estimated by examining the fraction of Non-CpGs that are methylated in Table 3.8. The overall difference in non-CpG (in both the CHG and CHH contexts) methylation between the two conditions is small (less than 1%), displays a trend opposite to that observed in methylation in the CpG context and is not statistically significant (p-value 0.36 and 0.71 for the CHG and CHH contexts respectively). The above in combination suggest that differential conversion efficiency is not the cause of the observed methylation trend.

Differential methylation was examined following an approach similar to that employed by the limma R package [Smyth, 2005]. The relationship of the standard deviation to the mean methylation was modelled using a spline function (Figure 3.15). The model predictions of standard deviation were used as input to multiple independent T-tests to identify differential methylation events. 1,641 differentially methylated CGIs were detected at the 0.05 p-value level of significance. However, none of these were significant after FDR multiple testing correction, even after stringent reduction of the number of CGIs examined on the basis of number of informative CpG per CGI. The top hits of these analysis are presented in Table A.2 of the Appendix. Given that the results of this analysis have not been corrected for multiple testing, a high number of false positives is expected and per locus validation of these results particularly important.

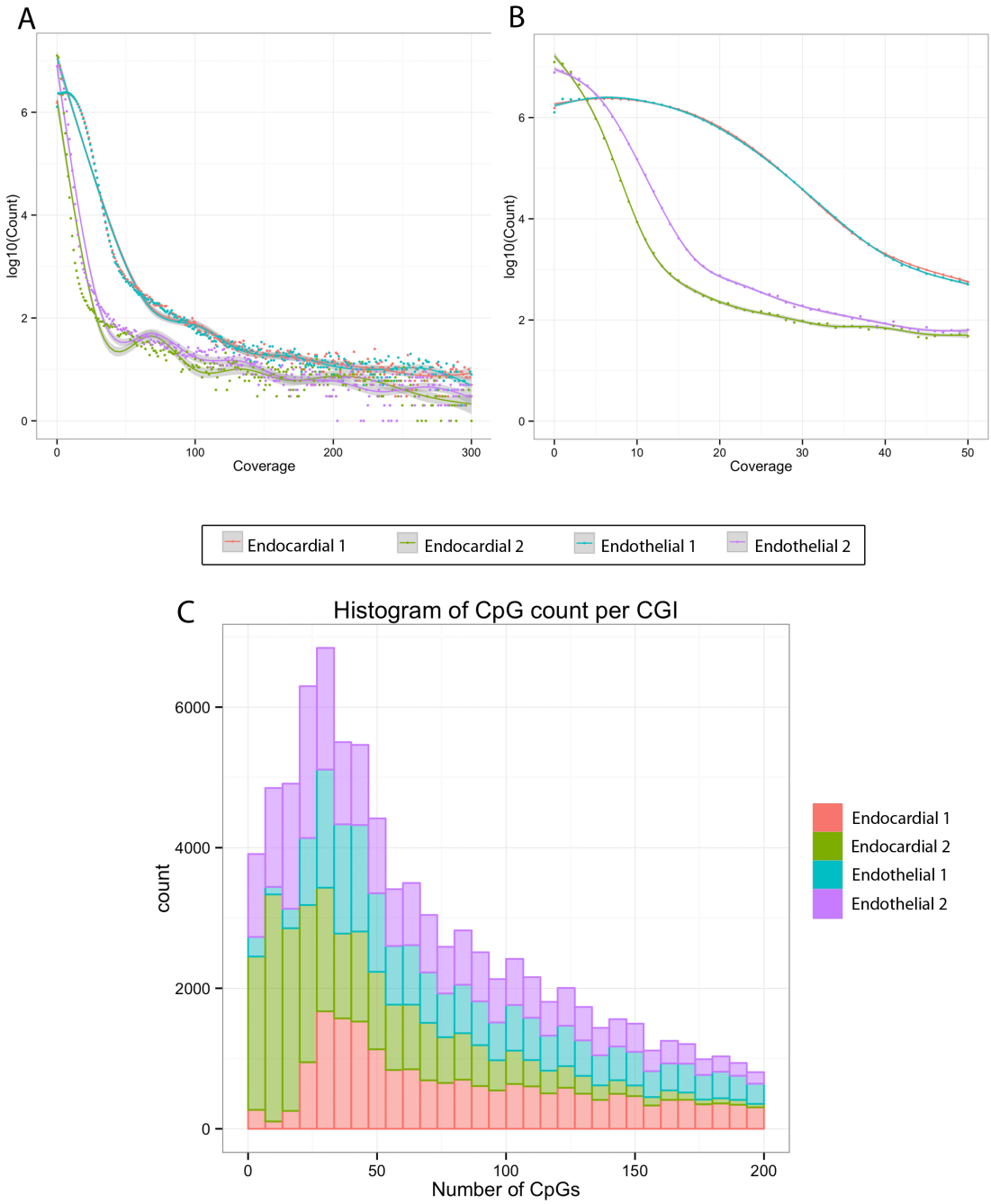


Figure 3.13: (A) Logarithmic scale histogram of CpG coverage in each sample. (B) Logarithmic scale histogram of CpG coverage in each sample for the coverage range 0 - 50. (C) Stacked histogram of informative CpGs per CGI for each sample.

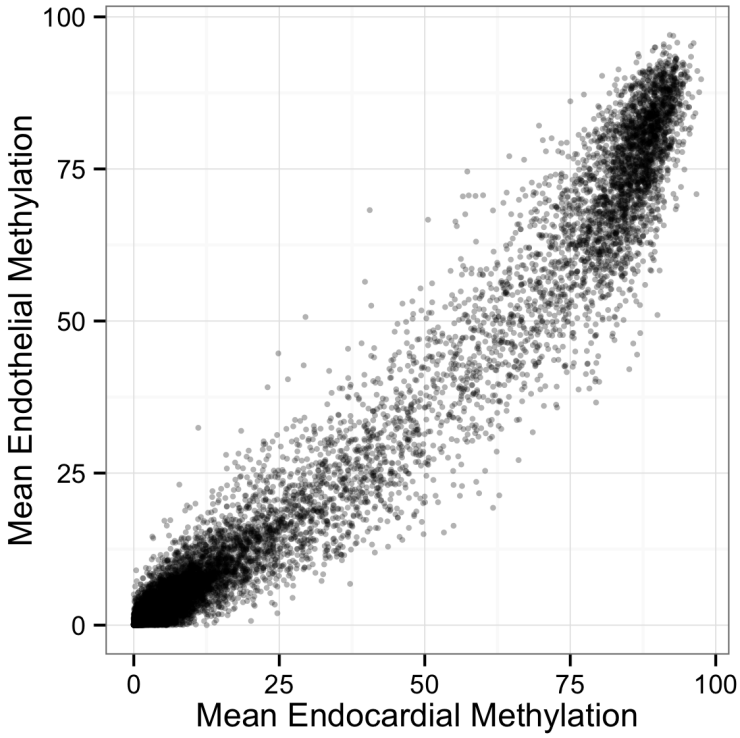


Figure 3.14: Scatter plot of mean methylation in endocardial and endothelial cells. The scatter plot suggests hypermethylation of endocardial DNA.

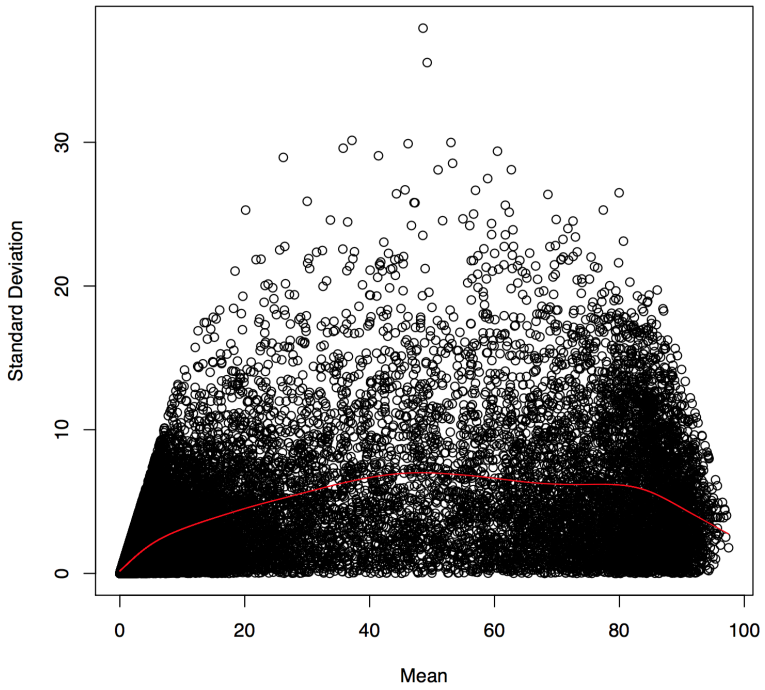


Figure 3.15: Relationship between standard deviation and mean methylation of CGIs. The relationship was modelled with a spline function (red line) and the predicted values were used for multiple independent T-tests to assess methylation differences.

Table 3.8: Percent methylated C's in each genomic context. Endocardial CpG methylation is higher than endothelial methylation, in contrast to that of other genomic contexts that do not recapitulate this trend, suggesting that this difference is not the result of differential conversion rates between the samples.

Sample	%me CpG	%me CHG	%me CHH
EC replicate 1	73.87%	0.53%	0.43%
EC replicate 2	76.20%	1.40%	1.40%
ET replicate 1	70.85%	0.53%	0.43%
ET replicate 2	74.50%	2.20%	2.30%
EC mean	**75.03%**	0.97%	0.92%
ET mean	72.68%	**1.36%**	**1.36%**

3.2.8 Genomic Context of Differentially Methylated CGIs

The genomic distribution of differentially methylated CGIs (at the p-value <0.05 significance level) was examined and compared with that of all CGIs examined (Figure 3.16). The distribution of the differentially methylated CGIs did not differ appreciably from that of all the genomic CGIs examined. Most noticeable was a depletion in distal intergenic regions and enrichment in Promoters and 5'UTRs, potentially signifying a regulatory role in gene expression of a subset of these CGIs.

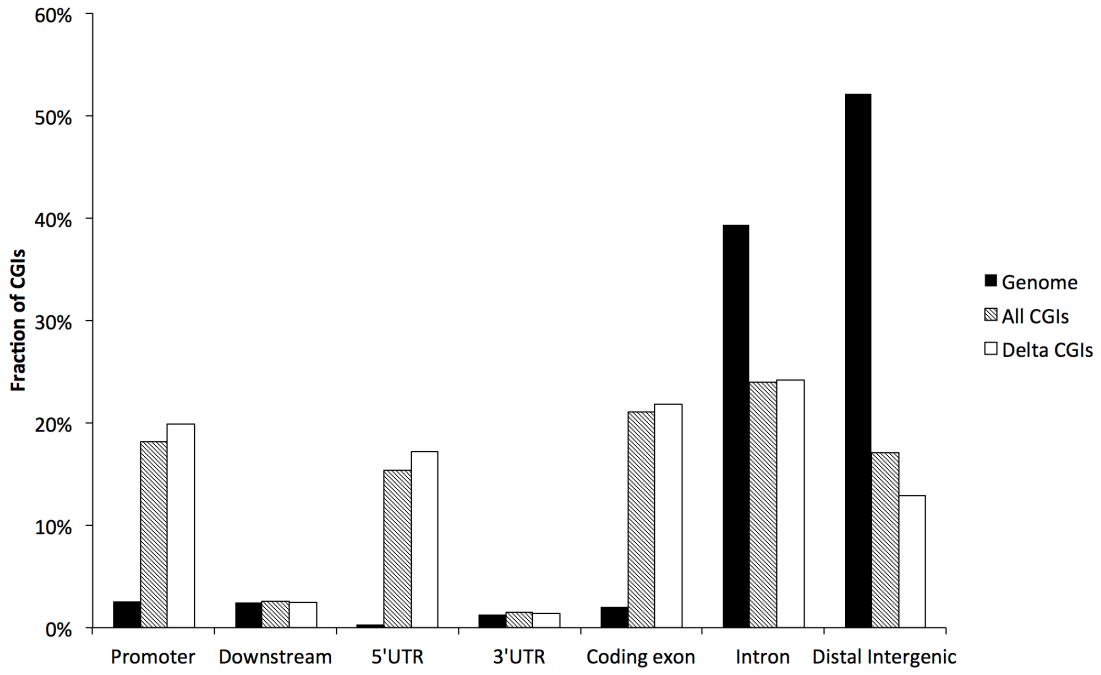


Figure 3.16: Genomic distribution of all examined and differentially methylated CGIs in comparison to genomic composition. Differentially methylated CGIs are underrepresented in distal intergenic regions and are enriched in promoters, 5' UTRs and coding exons.

3.2.9 Detection of Genome-wide Differential Methylation

Further to the per-CGI analysis described above differential methylation was detected on a genome-wide level using the BSmooth algorithm [Hansen et al., 2012]. BSmooth generates a smoothed methylation profile for each replicate independently by sharing methylation information between nearby CpGs. Identification of differentially methylated regions is subsequently performed by calculation of T-statistic and identification of genomic regions consistently displaying extreme T-statistic values.

In the context of this analysis, DMRs were called as significantly differentially methylated if the T-statistic was beyond the lower or upper 1% of its empirical distribution and furthermore the mean difference in methylation between the two samples exceeded 20% and the locus was supported by at least 20 informative CpGs. These cutoffs were more stringent that the values recommended by the BSmooth algorithm authors.

This analysis identified a total of 1,128 differentially methylated regions between the endocardium and the endothelium. Most (1,083/1,128) of these sites were hypermethylated in the endocardium (Table 3.9), consistent with the observations of the per-CGI analysis presented above. The size of the detected DMRs was in the range of 116 to 1,661 bps with a median size of 539 bps. Full results of this analysis can be found in Table A.3 of the Appendix. The five significant loci that displayed the maximal methylation differences between the two tissues are shown in Figure 3.19 as examples.

The following sections examine the genomic context of the DMRs identified here, their overlap with other epigenetic marks and the functional roles of genes in close proximity.

Table 3.9: Number of identified hyper- and hypo- methylated genome-wide differentially methylated genomic regions. The majority of the regions are hypermethylated consistent with previous observations in the course of this analysis.

Methylation Status in EC	DMR Count
Hypermethylated	1,083
Hypomethylated	45
Total	1,128

3.2.10 Genomic Context of Differentially Methylated Genomic Regions

The genomic distribution of the genomic DMRs identified above was examined and is presented in Figure 3.17 where it is compared to the proportion of the genome constituting each class of genomic sequence.

The majority (62%) of the detected DMRs are in the vicinity of coding regions, while 38% are in distal intergenic regions. Despite that, DMRs are underrepresented in intronic regions and distal intergenic locations in comparison to the overall genomic composition. Furthermore, DMRs are overrepresented in the vicinity of annotated genes and particularly in coding exons. This distribution suggests that the differential methylation may be linked to the differential regulation of expression of transcripts. The relationship between differential methylation and differential expression is examined later in this thesis.

In addition to examining the distribution of DMRs with respect to functional status of the genome, the distribution of DMRs was compared to the DNAse I, H3K36me3, H3K27me3, H3K27ac and H3K4me3 profiles from cardiac tissue obtained from the ENCODE project [The ENCODE Project Consortium, 2012]. Cardiac tissue constitutes a useful proxy for the epigenetic profiles of endocardium as both are tissues of mesodermal origin and developmentally related. However, its utility is limited by the considerable developmental distance between the two tissues. The results of this analysis are presented in Figure 3.18.

The overlap of all the marks examined was higher than expected by chance ($p < 0.001$ for all marks) as evaluated by means of permutation testing. Despite the considerable developmental distance between the endocardium and the adult heart, 508 of the 1,128 DMRs identified, overlapped a histone mark or a DNAse I hypersensitive site in heart, suggesting functional significance of the differentially methylated sequences. Overall, the examination of the genomic location of the identified DMRs suggests that at least some of these sites are functionally important and the methylation changes observed are biologically relevant.

Given that considerable overlap between the identified DMRs and transcribed genomic regions was observed in the above analysis, the specific overlap of the identified DMRs with known genes was examined. Of the 1,128 identified DMRs, 696 were found to directly overlap the genomic location of 381 unique genes. GO term overrepresentation analysis of the list of these genes revealed significant associations with terms related development processes such as “developmental process” and “anatomical structure morphogenesis” (Table 3.10). This finding, strongly suggests a functional role for these DMRs in development of the endocardium.

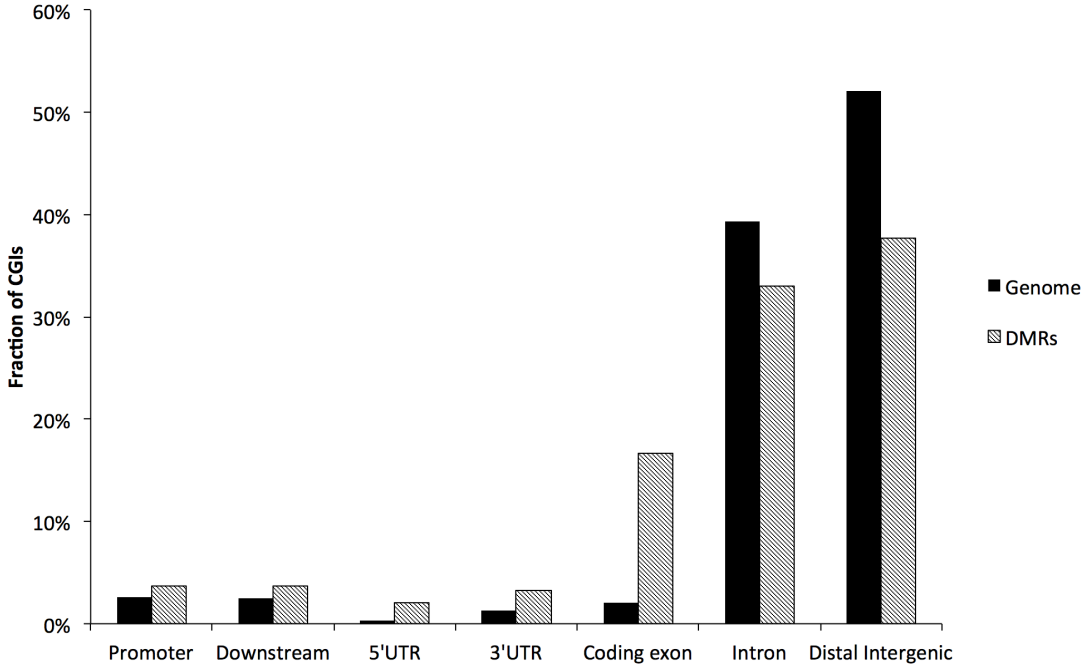


Figure 3.17: Genomic distribution of differentially methylated regions in comparison to genomic composition. DMRs are overrepresented in the vicinity of annotated genes, but not introns or distal intergenic regions.

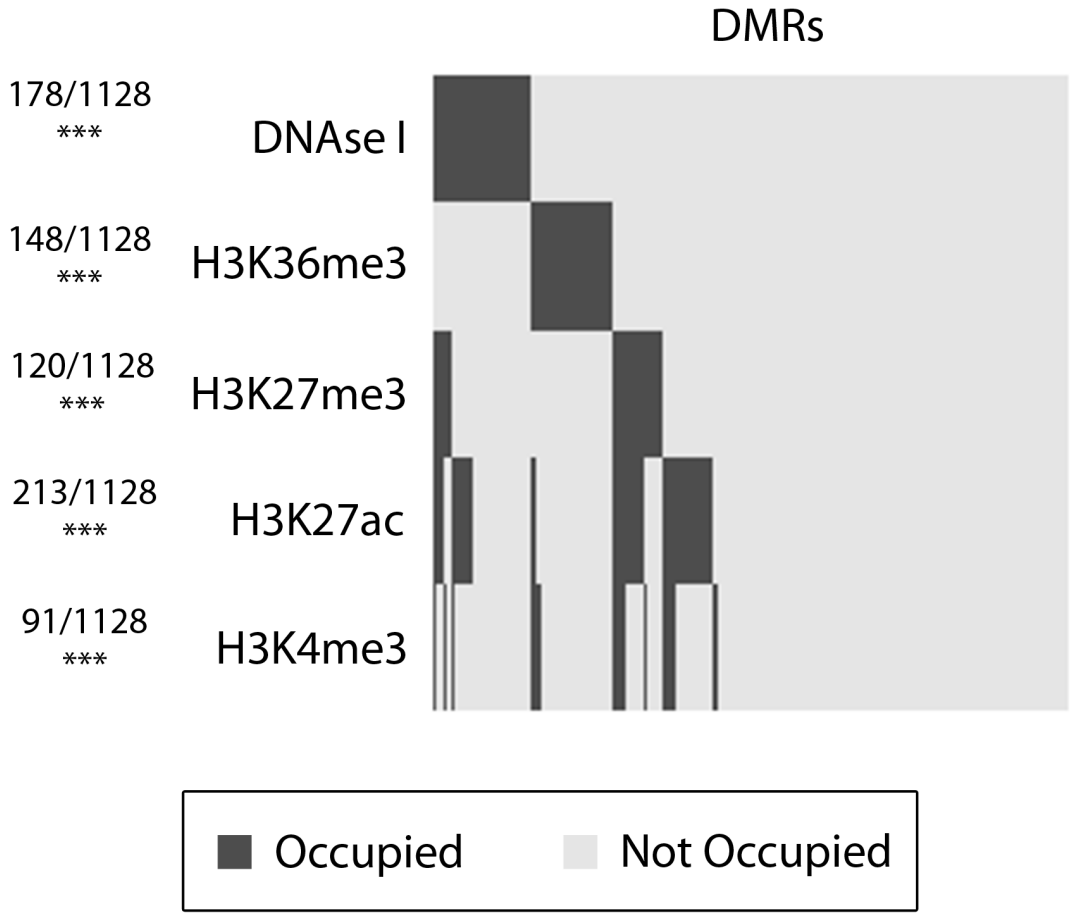


Figure 3.18: Heatmap of overlap of peaks as identified by the ENCODE project with the DMRs identified in this study. The numbers on the left signify the number of DMRs that were found to overlap each mark. The significance of the overlaps was assessed by means of permutation testing ($p < 0.001$ in all cases, denoted by ***). Overall, 508 of the 1,128 (45%) DMRs were found to overlap at least one mark suggesting that they are functionally significant.

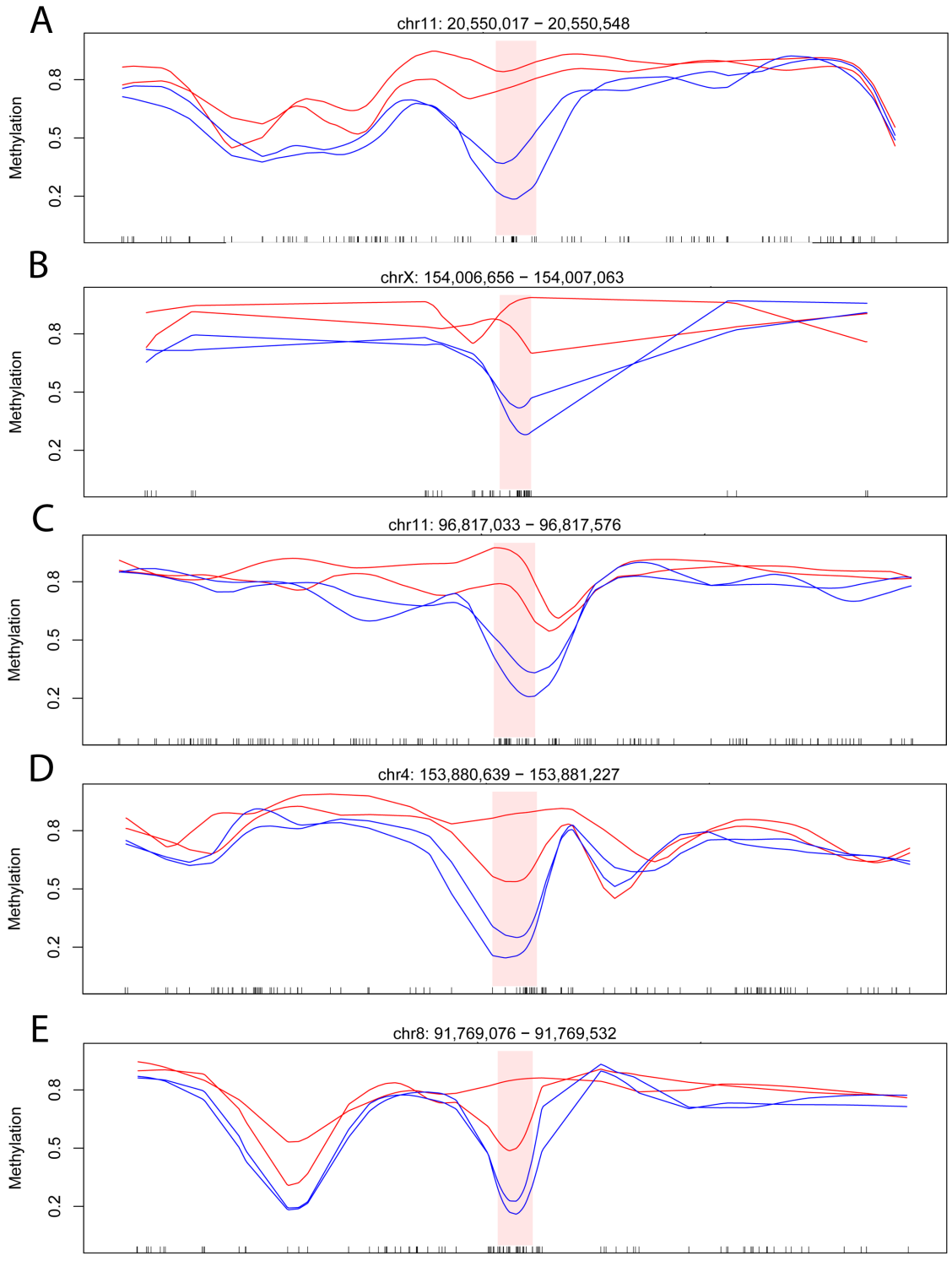


Figure 3.19: Visualisation of the methylation profile of the top five most significant differentially methylated genomic locations (pink boxes) and 5 kb flanking regions. Individual CpGs are denoted as notches at the bottom of the plots. Endocardial replicates are displayed red and endothelial in blue.

Table 3.10: Significantly overrepresented Molecular Function GO Terms in the set of genes directly overlapping identified DMRs between the endocardium and the endothelium. The terms reveal a connection of the identified DMRs with developmental processes, strongly suggesting a functional role for these sites in the development of the endocardium.

GO Term ID	Description	P-value	FDR q-value
GO:0044767	single-organism developmental process	1.46E-12	1.67E-08
GO:0032502	developmental process	3.76E-12	2.15E-08
GO:0009653	anatomical structure morphogenesis	1.81E-09	5.19E-06
GO:0048856	anatomical structure development	1.38E-09	5.26E-06
GO:0048869	cellular developmental process	5.42E-09	1.24E-05
GO:0048858	cell projection morphogenesis	1.25E-07	2.39E-04
GO:0048812	neuron projection morphogenesis	2.78E-07	4.55E-04
GO:0007275	multicellular organismal development	3.64E-07	4.64E-04
GO:0032990	cell part morphogenesis	3.39E-07	4.85E-04
GO:0044763	single-organism cellular process	5.04E-07	5.78E-04
GO:0032989	cellular component morphogenesis	1.09E-06	1.14E-03
GO:0030030	cell projection organization	2.63E-06	2.52E-03
GO:0016043	cellular component organization	4.52E-06	3.46E-03
GO:0009987	cellular process	4.03E-06	3.55E-03
GO:0048513	organ development	4.47E-06	3.66E-03
GO:0048646	anatomical structure formation involved in morphogenesis	6.04E-06	4.08E-03
GO:0030154	cell differentiation	6.02E-06	4.31E-03
GO:0071840	cellular component organization or biogenesis	9.89E-06	6.30E-03
GO:0035108	limb morphogenesis	2.73E-05	1.56E-02
GO:0035107	appendage morphogenesis	2.73E-05	1.65E-02
GO:0008347	glial cell migration	3.34E-05	1.74E-02
GO:0002040	sprouting angiogenesis	3.34E-05	1.83E-02
GO:0044260	cellular macromolecule metabolic process	3.90E-05	1.94E-02

GO:0051093	negative regulation of developmental process	4.06E-05	1.94E-02
GO:0048519	negative regulation of biological process	4.68E-05	2.07E-02
GO:0010604	positive regulation of macromolecule metabolic process	4.53E-05	2.08E-02
GO:0043170	macromolecule metabolic process	5.18E-05	2.20E-02
GO:2000026	regulation of multicellular organismal development	8.13E-05	3.33E-02
GO:0002175	protein localization to paranode region of axon	9.70E-05	3.84E-02
GO:0008283	cell proliferation	1.06E-04	4.06E-02
GO:0044237	cellular metabolic process	1.12E-04	4.16E-02
GO:0045944	positive regulation of transcription from RNA polymerase II promoter	1.44E-04	5.17E-02
GO:0050793	regulation of developmental process	1.69E-04	5.55E-02
GO:0009893	positive regulation of metabolic process	1.69E-04	5.70E-02
GO:0010648	negative regulation of cell communication	1.94E-04	5.71E-02
GO:0035113	embryonic appendage morphogenesis	1.92E-04	5.79E-02
GO:0010721	negative regulation of cell development	1.67E-04	5.81E-02
GO:0006807	nitrogen compound metabolic process	2.05E-04	5.88E-02
GO:1901343	negative regulation of vasculature development	1.92E-04	5.95E-02

Table 3.11: RNA Samples, cell types and description of DNA samples used for RNA-seq library preparation. The letters A-H are used throughout to refer to the particular libraries.

Sample ID	Sample Name	Cell Type	Description
A	s1_ec	EC	Endocardial Replicate 1
B	s1_et	ET	Endothelial Replicate 1
C	s2_ec	EC	Endocardial Replicate 2
D	s2_et	ET	Endothelial Replicate 2
E	s4_ec	EC	Endocardial Replicate 3
F	s4_et	ET	Endothelial Replicate 3
G	old_ec	EC	Endocardial Replicate 4
H	old_et	ET	Endothelial Replicate 4

3.3 Transcriptome Analysis of Endocardial and Endothelial Cells from Embryoid Bodies

3.3.1 Initial Reanalysis of Preexisting mRNA-seq Data

At the commencement of this project externally generated RNA-seq data were available for endocardial and endothelial cells. Initial data quality assessment with FastQC demonstrated a very high duplication rate, estimated in excess of 70% and generally low quality of the reverse reads. First reads were de-duplicated without reference to the genomic alignment and the remaining data were processed with the Tuxedo suite (as outlined in the next sections), but did not yield any significant hits and further work on these data was abandoned.

3.3.2 Experimental Design

RNA-seq was repeated on endocardial and endothelial cells in its entirety in the context of this project. Four pseudo-biological replicates from independent embryoid body culture pools were sequenced for each cell type. Four replicates were employed as past experience with these cell types suggested that any effect sizes would be small and therefore a higher number of replicates was desired to detect these changes to a high significance level (Prof. Baldwin, personal communication).

3.3.3 Library Preparation and Sequencing Results

Frozen FACS sorted cell samples were shipped from the Baldwin laboratory. RNA was extracted, quantified and its quality determined as outlined in Section 2.7.1 and before

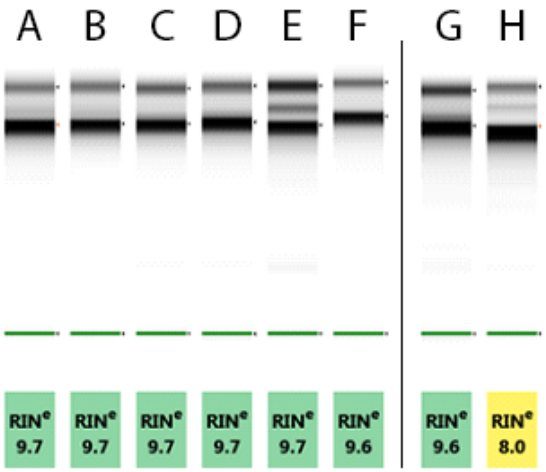


Figure 3.20: Gel representation of TapeStation data for RNA-seq input RNA. The quality of all samples is equal to or exceeds the minimum of 8.0 RIN.

libraries were prepared (Section 2.7.4). Identifiers, sample names, and sample description can be found in Table 3.11.

The samples were quantified using the Qubit instrument, results of the quantification are shown in Table 3.12, the total amount of RNA was in all cases above the minimum 0.1 µg required. The quality of the extracted RNA was assessed using the Tapestation RNA High Sensitivity assay (Figure 3.20). All samples had a RIN of 8.0 or higher and were suitable for library preparation.

Table 3.12: RNA sample quantification and estimated volume and calculation of total RNA. RNA was above minimum (0.1 µg) required for library preparation.

Sample	Readings (µg/mL)			Volume (µL)	Total RNA (µg)
	1st	2nd	Mean		
A	9.74	9.56	9.65	46	0.44
B	9.24	9.27	9.26	46	0.42
C	10.10	10.20	10.15	46	0.43
D	8.47	8.58	8.53	46	0.39
E	24.10	23.60	23.85	46	1.10
F	4.10	4.90	4.50	46	0.21
G	13.00	12.80	12.90	46	0.59
H	15.70	14.60	15.2	46	0.70

Directional poly(A) selected libraries were prepared as discussed in Section 2.7.4. The fragment size of the libraries was estimated using the Agilent Tapestation instrument (see Figure 3.21). Due to a fault of the reagent batch provided, accurate size estimation was not possible and insert sizes were, by comparison to past runs and the expected insert

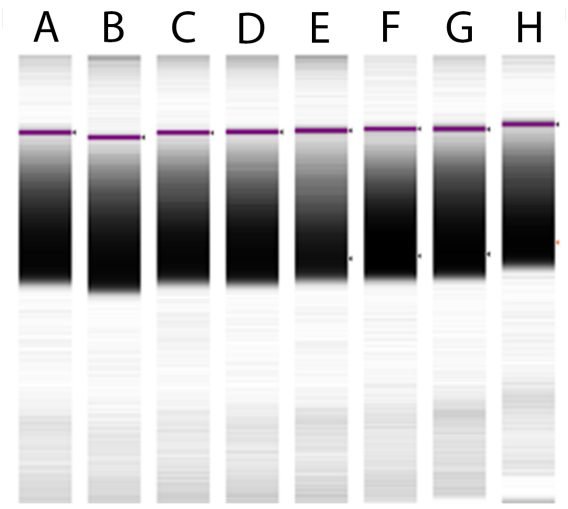


Figure 3.21: Gel representation of Tapestation fragment size distribution data used for RNA-seq library size estimation. Individual libraries are shown as separate lanes (A-H). The size distribution of DNA fragments in all libraries was similar and consistent with an insert size of 300 bp. Marker information was not complete and accurate size estimation was not possible.

size was estimated to 300 bps. This value was selected to be higher than the expected median fragment size, as use of a lower fragment size than that of the sample would lead to more concentrated libraries that would overload the flowcell lane and yield no useful data.

Libraries were diluted 1:1000 and quantified via qPCR (see Table 3.13), normalised to 10 nM and subsequently pooled. The pooled libraries was sequenced on a single lane of a Illumina®HiSeq 2000 instrument. Addition of phiX control DNA library, cluster generation and sequencing was performed by the BRC Sequencing Facility. The total cluster count of 83 million (see Table 3.14) was low compared to the optimal instrument run, which produces in excess of 100 million cluster counts. The lower number of clusters was however offset by the very high quality of reads and no read removal was required during quality control, as outlined in the next sections.

3.3.4 Bioinformatic Processing

The data produced during sequencing were bioinformatically processed to identify differentially regulated genes between the two cell types. Processing consisted of basecalling, assessment of the quality of the obtained data, filtering of the data based on the quality control, alignment to the transcriptome and genome, building of a reference annotation and differential expression analysis. In contrast to the BS-seq analysis outlined above phiX DNA removal was not explicitly performed as the individual libraries were barcoded

Table 3.13: RNA-seq library qPCR quantification and dilution calculations. The correction factor corrects for the insert size difference from the insert size fragment of the standard reference library and is used to calculate the diluted corrected library concentration. The final dilution volume is the final volume in which 1 μL of the original library must be diluted in to obtain a 10 nM library.

ID	[Diluted Library] (pM)	Fragment Size(bp)	Correction Factor	[Diluted Corrected Library] (pM)	[Undiluted Corrected Library] (pM)	Final dilution Volume (μL)
A	618	300	1.51	931	931	466
B	1028	300	1.51	1549	1549	774
C	1048	300	1.51	1579	1579	789
D	879	300	1.51	1324	1324	662
E	691	300	1.51	1041	1041	521
F	672	300	1.51	1012	1012	506
G	556	300	1.51	838	838	419
H	641	300	1.51	966	966	483

Table 3.14: Adaptor identifiers, adaptor sequences and Raw Cluster Counts for the mRNA-seq libraries. The total cluster count was lower than expected from a HiSeq lane, but sufficient for differential gene expression analysis.

Sample ID	Adaptor ID	Adaptor Sequence	Raw Cluster Count
A	AR005	ACAGTG(A)	9,293,011
B	AR006	GCCAAT(A)	12,422,156
C	AR012	CTTGTA(A)	11,140,829
D	AR019	GTGAAA(C)	7,837,534
E	AR002	CGATGT(A)	10,553,367
F	AR004	TGACCA(A)	11,097,369
G	AR007	CAGATC(A)	10,035,665
H	AR013	AGTCAA(C)	10,088,855
		Total	83,468,813

and phiX reads did not carry barcodes. These steps are outlined in more detail in the following sections.

Quality Control

Basecalling was performed with Casava as outlined in section 2.7.5. Quality control was performed with the FastQC toolkit. A representative output example is shown in Figure 3.22; output from FastQC runs on the entire dataset were visually examined. The read quality was judged to be adequate on the basis of the base quality per position plot (Fig 3.22 A) and reads were not removed or trimmed on the basis of this score.

Unexpectedly, the base pair composition at the 5' end of all reads (both forward and reverse) was not uniform as expected (Fig 3.22 B). The source of this characteristic was not identified but it is hypothesised to be the result of sequence-dependent fragmentation, due to fragmentation by chemical means being employed (as opposed to sonication). In addition, the presence of self-ligated adaptors could be the source of this contamination. The presence of a k-mer poly(A) peak further suggest that the fragmentation is not random (Fig 3.22 F).

On this basis the first 20 base-pairs of each read were removed and not examined in the subsequent steps. This was done to minimise mapping bias from these sequences. Given the long paired-end sequencing employed and the fact that the actual read sequence is of no importance in RNA-seq – other than to ensure correct mapping of the reads – this is expected to have no adverse effect in subsequent data analysis.

DNA duplication rates as identified by FastQC were very high (in excess of 50%) and were initially a cause for concern. It must be noted that FastQC utilises a hash table to estimate duplication rates and only tracks the first 50 bps of the first two hundred thousand sequences; furthermore it does not take into account the paired nature of the reads. The reported high duplication at this stage was ignored and duplication estimation and removal was performed after alignment taking into account both ends of each fragment. This revealed a much lower duplication rate.

Alignment

Alignment was performed using Tophat 2 and Bowtie 2 as the underlying aligner against the mm9 version of the mouse genome as outlined in Section 2.7.5. Alignment was performed against the annotated transcriptome and any unaligned reads were subsequently

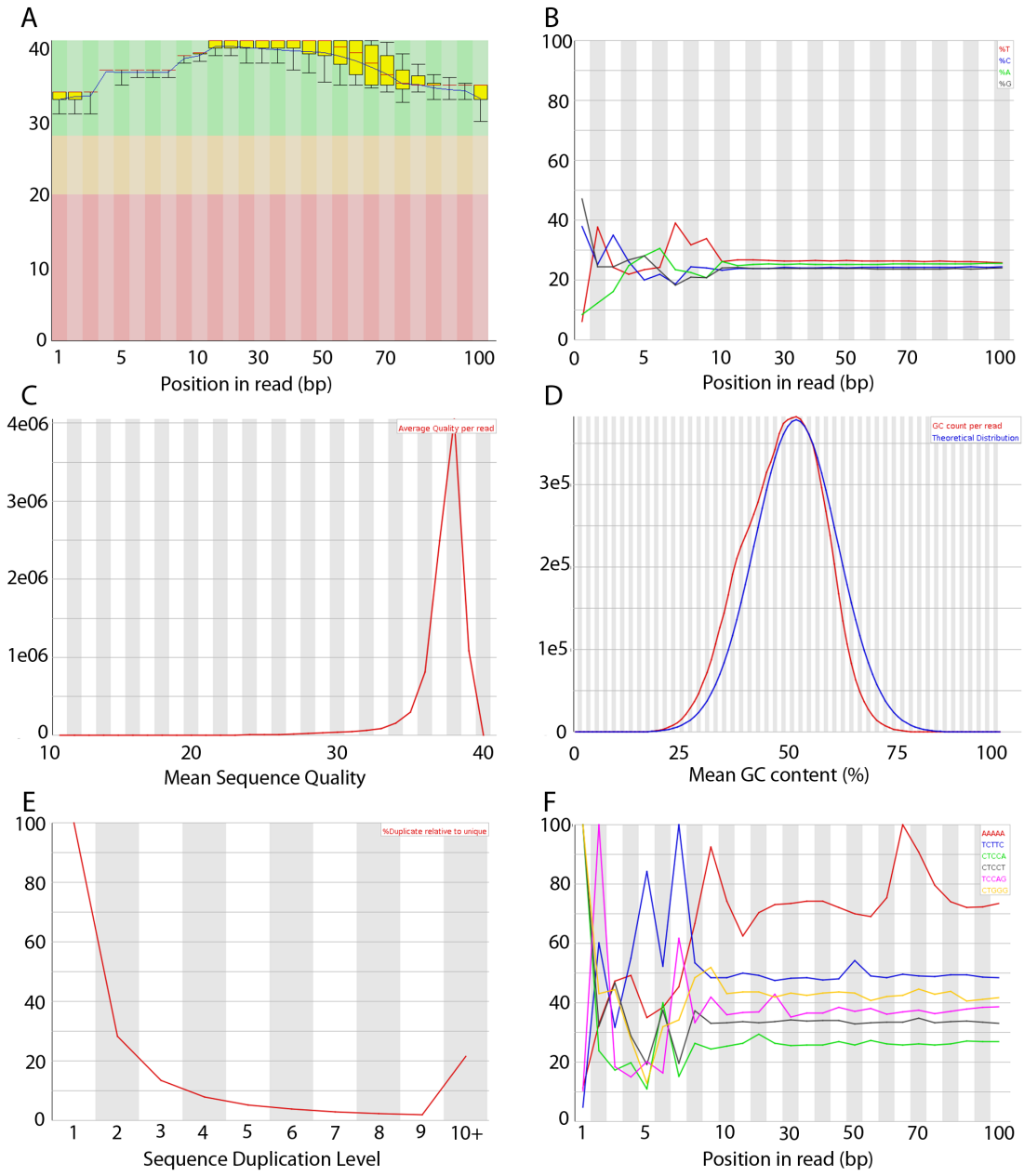


Figure 3.22: Initial quality control plots of RNA-seq sample A first reads; these plots are representative of other samples. (A) Phred scaled quality score box plot as a function of read position reveals very high quality of all reads (B) Base incorporation percentage as a function of read position shows variable composition at the 5' end but stable throughout the read, this could be attributed to non-random fragmentation. (C) Average quality per read distribution plot, shows very high quality of all reads with a short left-hand tail. (D) GC% distribution closely matches the expected distribution. (E) Duplicate distribution reveals very high duplication rate, this plot is not representative of the true duplication rate (see main text). (F) k-mer content as a function of read position is highly uniform after the first 10 bp, with the exception of poly(A) overrepresentation, which is consistent with mRNA-seq.

Table 3.15: Counts of aligned and concordantly aligned unique read pairs, show a very low percentage of discordant read pairs and a high overall unique alignment rate.

Sample	Input	Aligned	Discordant	Condordant Unique
A	9,293,011	8,578,317	1.6%	7,956,505 (86%)
B	13,422,156	12,463,769	1.4%	11,636,058 (87%)
C	11,140,829	10,443,478	1.3%	9,835,121 (88%)
D	7,837,534	7,232,682	1.4%	6,805,494 (87%)
E	10,553,367	9,870,866	1.2%	9,378,143 (89%)
F	11,097,396	10,307,804	2.0%	9,465,814 (85%)
G	10,035,665	9,376,667	1.2%	8,832,287 (88%)
H	10,088,855	9,339,070	1.4%	8,807,181 (87%)
			Mean	87%

aligned to the genomic sequence to identify novel transcripts. As shown in Table 3.15, approximately 87% of all read pairs were uniquely and concordantly aligned to the transcriptome, which is sufficient for the purposes of this investigation and above the expected value. As the number of non-condordant reads was very low and did not exceed 2.0% in any sample these reads were not removed.

Duplicate Read Identification and Removal

As aforementioned, initial duplication estimation suggested a very high duplication rate. Duplicate identification and removal was performed with the Picard toolset as outlined in Section 2.7.5. This revealed a much lower duplication rate of approximately 10% (Figure 3.23).

Estimation of the duplicate rate with the Picard software package is much more robust than estimation with FastQC, as it takes into account the end position of each sequenced fragment and can therefore distinguish between fragments that start at the same position due to chance. Overestimation of the duplication rate by FastQC highlights the difficulties of accurately removing duplicates when aligning against a small reference (in this case the transcriptome), as identical reads become more frequent due to chance and are not PCR amplification artefacts.

Annotation Building

A transcriptome annotation was built independently for every sample using the `cufflinks` program. Individual annotations were merged using `cuffmerge` as described in Section 2.7.5. The merged annotation contains 25,263 genes of which 23,221 (91.1%) are present

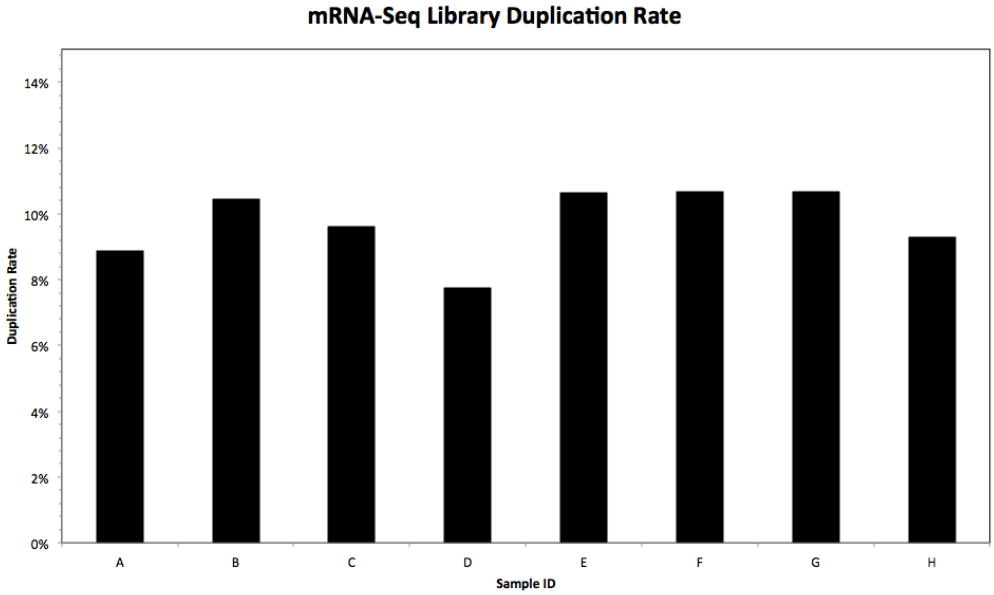


Figure 3.23: Duplication rate per library for mRNA-seq libraries as identified post-alignment by the Picard toolkit, were low and did not exceed 12% for any of the samples. Duplication rates calculated with Picard were in disagreement with duplication rates calculated with FastQC (see main text for details).

in the reference annotation of 23,366 transcripts.

This suggests that 2,042 novel transcripts have been identified, whereas 145 known transcripts have not been included in the annotation (Figure 3.24). The 145 genes not included in the annotation were identified and a manual inspection of the mapped reads and merged annotation was performed in approximately 20 of these loci at random. In all cases the merged annotation included a transcript of the correct structure at the location, but the transcript was not correctly mapped to the common gene identifier. This suggests that a programmatic error in `cuffmerge` prevents correct annotation of a small portion of the genes in the final annotation, but does not result in their exclusion from the analysis. This is therefore not expected to adversely affect subsequent analysis.

Confirmation of Expression Pattern of NFATc1

The read mapping pattern of NFATc1, the marker used for separation of the two examined cell populations via FACS, was examined using the UCSC genome browser prior to further analysis (Figure 3.25). A considerably higher number of reads mapped to *NFATc1* in endocardial samples. Given the approximately equal number of reads in every library, (see Table 3.15) the excess of reads in the endocardial samples suggested upregulation of NFATc1 in the endocardium and confirmed that no sample mislabelling had occurred.

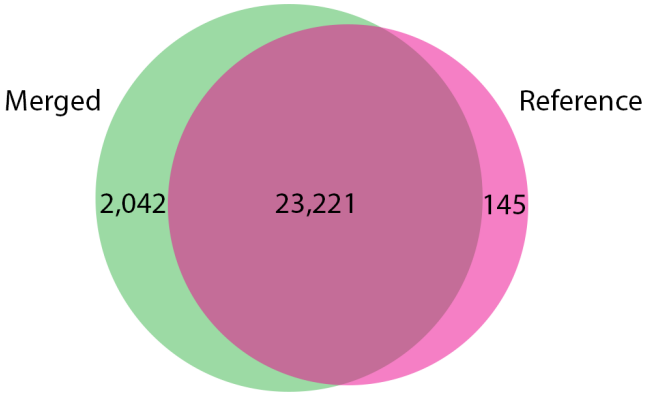


Figure 3.24: Comparison of Reference UCSC annotation from refFlat with annotation built from mRNA-seq on endocardial and endothelial cells, reveals that the overwhelming majority of annotated genes were detected in at least one cell type and 7,384 novel transcripts were discovered.

3.3.5 Differential Expression of Genes and Transcripts

Differential expression analysis was performed with `cuffdiff` as outlined in Section 2.7.6. This algorithm was selected for the differential expression analysis of this dataset after comparison of the results between `cuffdiff`, `limma` [Smyth, 2005] and `DESeq` [Anders and Huber, 2010]. The comparison revealed that `cuffdiff` was the only algorithm that successfully identified the NFATc1 gene, the known marker on which the two cell populations were selected as differentially regulated between the two cell types.

The FPKM distribution of individual libraries was assessed (Figure 3.26 A). Libraries EC1 and EC4 displayed considerable deviation from the FPKM distribution of other samples. This was however not recapitulated in the FPKM histogram plot (Figure 3.26 B). It is of interest to note that libraries EC1 and EC4, displayed the highest number of missing values for genes in the annotation and this resulted in the boxplot being skewed, whereas the histograms, which do not report missing values, are not. When genes with missing values were manually excluded, the distribution of all samples was highly similar to each other (Figure 3.27) and consistent with a stable FPKM range of values.

The CV as function of FPKM is reported in Figure 3.26 C. The CV is as expected higher for both samples at low expression values and smaller at higher expression values. The CV is also consistently higher for endothelial samples. This is expected because the endothelial samples represent a more diverse and heterogenous population compared to

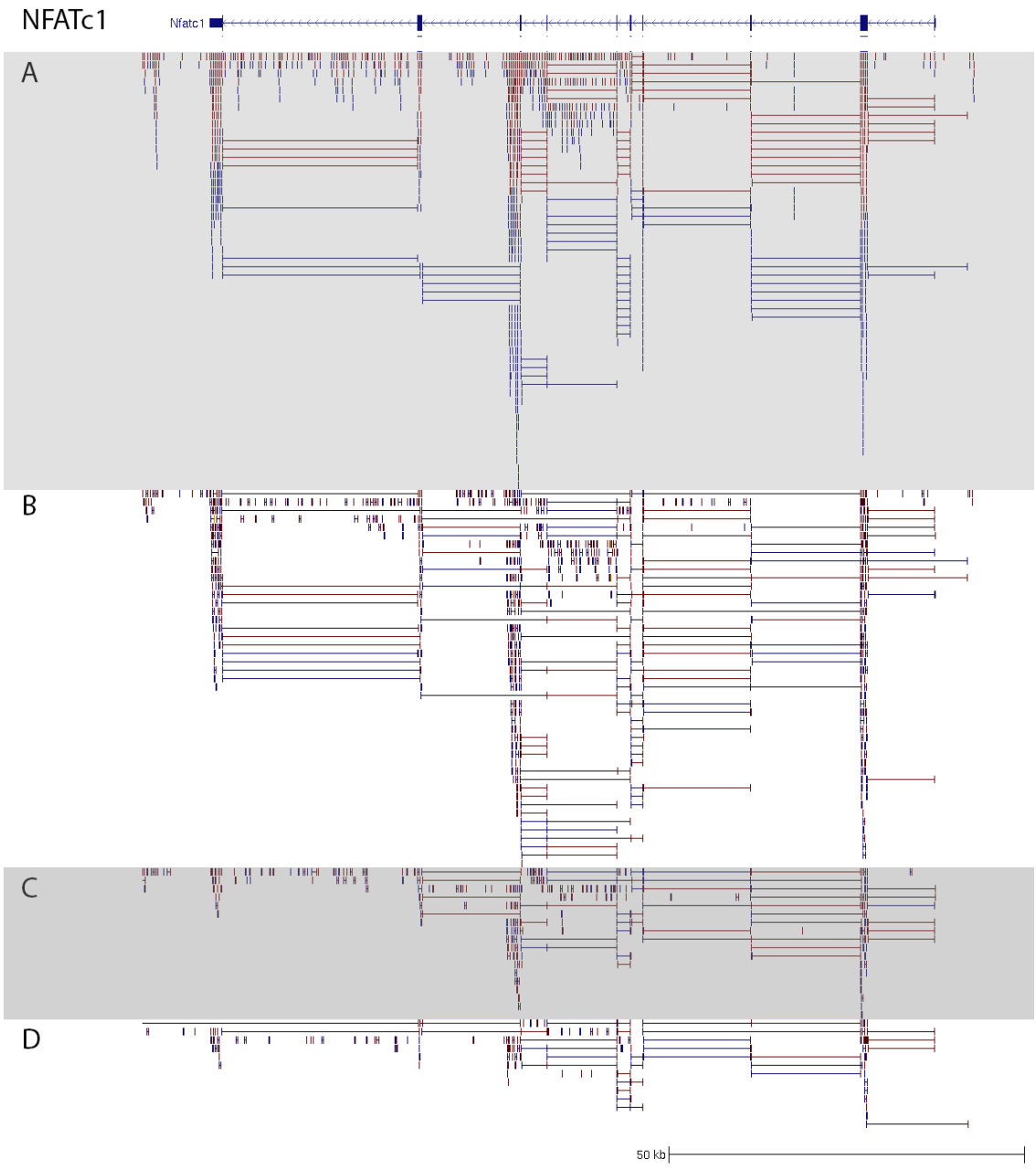


Figure 3.25: Raw mapped reads at the NFATc1 locus from the mRNA-seq experiment. (A,B) Reads from two replicates of the endocardial samples. (C,D) Reads from two replicates of the endothelial samples. Visual inspection confirms that samples are correctly labelled as NFATc1 is over-expressed in endocardial samples. All four replicates of each sample were examined during the analysis.

the more restrictive endocardial cell population.

The analysis identified 711 differentially expressed genes between endocardial and endothelial cells, of which 299 were upregulated in the endocardium and 411 downregulated in the endocardium. A plot of the negative log-transformed p-values against the log transformed fold change (Volcano plot) is shown in Figure 3.28. Note that the `cuffdiff` program utilises an empirically generated distribution to calculate p-values and in the interest of processing speed p-values are capped to minimum of 5 10^{-5}. This exceeds the significance threshold.

The top hits, by fold-change are presented in Tables 3.16 and 3.17 for genes upregulated in the endocardium and endothelium respectively. The complete list of differentially regulated genes can be found in the Appendix Table A.4.

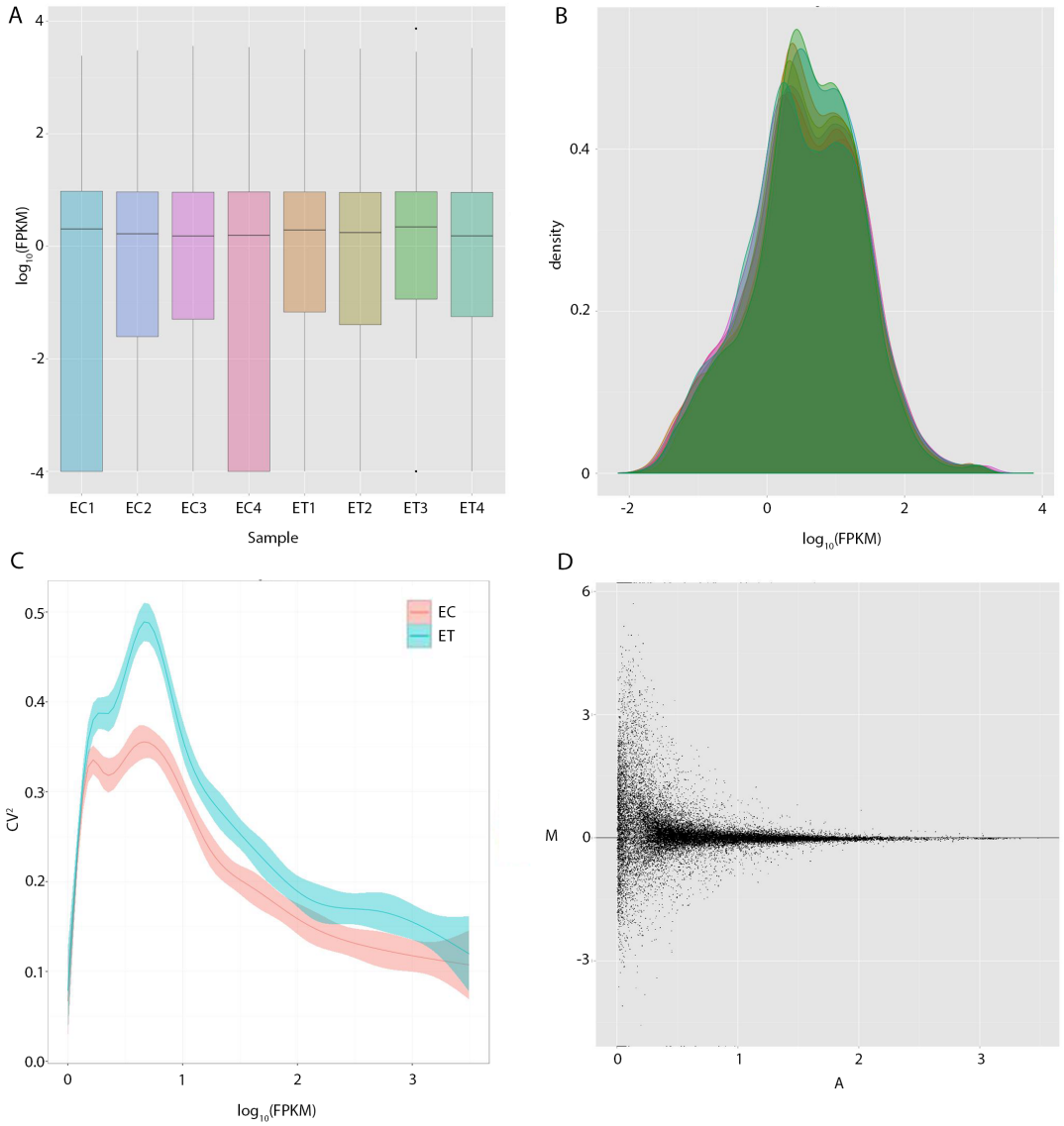


Figure 3.26: (A) Boxplot of log transformed FPKM values for individual libraries. Libraries EC1 and EC4, the libraries with the lowest read counts, show an unusual distribution due to missing values, see text page 129 for details. (B) Per-replicate FPKM distribution, samples EC1 and EC4 do not show an unusual distribution, suggesting that the unusual pattern in panel A is due to inclusion of these values as noughts. (C) Coefficient of variation as a function of log transformed FPKM values. Consistent with a more heterogenous population, endothelial cells show a consistently higher coefficient of variance. (D) Plot of mean expression in both cell types (A, x-axis) vs difference of means of each sample (M) shows no systematic trends in expression differences as a function of absolute expression.

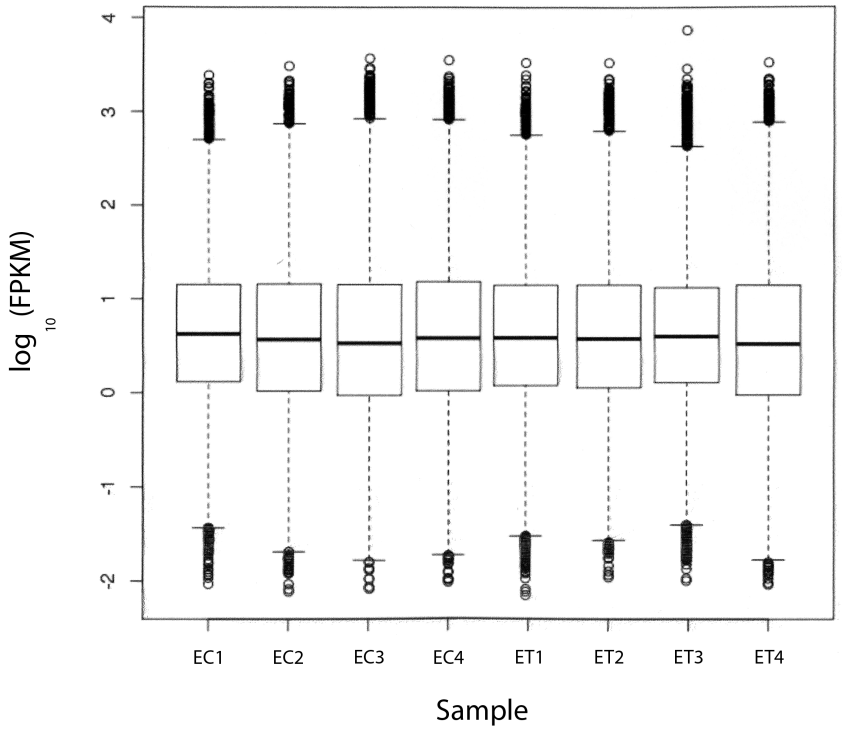


Figure 3.27: Boxplot of $\log_{10}$ transformed FPKM values for endocardial and endothelial samples after discarding missing values, on a per sample basis. The distributions are highly similar demonstrating that the sample specific differences observed can be attributed to missing data.

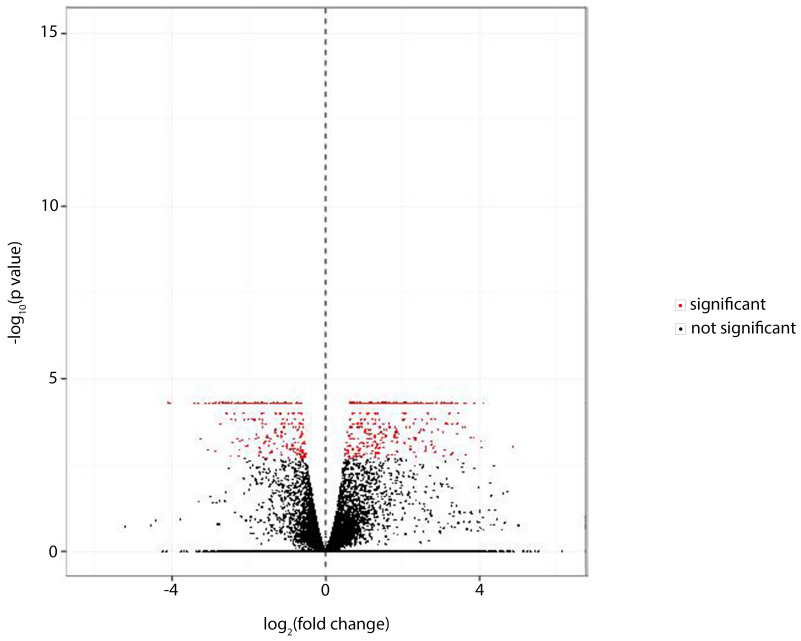


Figure 3.28: Volcano plot of differential expression between endocardial and endothelial cells. Significant hits appear in red. The p-value axis are capped, due to the way that `cuffdiff` generates p-values.

Table 3.16: Top 50 significantly upregulated genes in endocardial cells by fold change.

Gene	Locus	EC expression	ET expression	p-value	q-value	Fold Change
Oit3	chr10:58885707-58904527	5.20	0.30	5.00E-05	2.50E-03	17.15
Slc32a1	chr2:158436493-158441483	5.13	0.31	5.00E-05	2.50E-03	16.56
Gad2	chr2:22477846-22549397	4.57	0.43	5.00E-05	2.50E-03	10.68
Ptprb	chr10:115738429-115826594	2.53	0.25	5.00E-05	2.50E-03	9.93
-	chr15:72333348-72335718	2.09	0.22	5.50E-04	1.78E-02	9.60
Sost	chr11:101823771-101828329	4.49	0.51	5.00E-05	2.50E-03	8.77
Ccm2l	chr2:152891690-152907471	2.85	0.34	1.15E-03	3.19E-02	8.30
Cdh5	chr8:106625524-106668402	37.48	4.59	5.00E-05	2.50E-03	8.17
Sox7	chr14:64562542-64569569	10.68	1.39	5.00E-05	2.50E-03	7.67
Erg	chr16:95581810-95751972	11.47	1.51	5.00E-05	2.50E-03	7.58
Nos3	chr5:23870636-23897961	6.88	0.94	5.00E-05	2.50E-03	7.35
Grap	chr11:61466822-61486279	6.70	0.91	5.00E-05	2.50E-03	7.33
Lyve1	chr7:117994120-118006467	1.68	0.23	1.25E-03	3.41E-02	7.31
Dusp2	chr2:127161894-127164113	7.11	0.98	5.00E-05	2.50E-03	7.24
Rasip1	chr7:52882906-52894462	12.56	1.84	5.00E-05	2.50E-03	6.82
Eltd1	chr3:151100845-151208045	3.66	0.54	5.00E-05	2.50E-03	6.80
Skap1	chr11:96325904-96620936	5.66	0.84	5.00E-05	2.50E-03	6.78
Gimap4	chr6:48634576-48642061	2.02	0.30	2.00E-04	7.94E-03	6.69
Samsn1	chr16:75859038-75909511	4.03	0.62	5.00E-05	2.50E-03	6.48
Ushbp1	chr8:73908172-73919704	2.22	0.35	5.00E-05	2.50E-03	6.33
Tie1	chr4:118143795-118162454	40.45	6.62	5.00E-05	2.50E-03	6.11
Cd93	chr2:148262386-148269271	13.66	2.26	5.00E-05	2.50E-03	6.05
Kcne3	chr7:107325179-107333379	8.80	1.46	5.00E-05	2.50E-03	6.01
-	chr15:72337445-72342362	1.73	0.29	1.00E-04	4.56E-03	5.97
-	chr16:95580786-95581555	10.43	1.76	5.00E-05	2.50E-03	5.94
Adra2a	chr19:54119671-54123472	3.92	0.67	5.00E-05	2.50E-03	5.88
Icam2	chr11:106238969-106243955	13.43	2.29	5.00E-05	2.50E-03	5.87

Gimap6	chr6:48651581-48658243	6.43	1.11	5.00E-05	2.50E-03	5.80
Emcn	chr3:137004041-137094033	7.16	1.24	5.00E-05	2.50E-03	5.77
Ikzf1	chr11:11586215-11672929	4.18	0.74	5.00E-05	2.50E-03	5.65
Esam	chr9:37335673-37345904	3.97	0.72	1.70E-03	4.33E-02	5.54
Egfl7	chr2:26436575-26448202	12.09	2.19	5.00E-05	2.50E-03	5.52
Myct1	chr10:4739751-4752813	2.78	0.51	1.50E-04	6.36E-03	5.49
Myzap	chr9:71352153-71440167	9.76	1.79	5.00E-05	2.50E-03	5.46
Esam	chr9:37335673-37345904	27.65	5.10	5.00E-05	2.50E-03	5.42
Tspan8	chr10:115254339-115286949	2.12	0.39	7.00E-04	2.15E-02	5.41
Gfi1b	chr2:28464969-28477502	8.33	1.54	5.00E-05	2.50E-03	5.40
Abi3	chr11:95685813-95707045	4.32	0.83	1.00E-04	4.56E-03	5.19
Pcdh12	chr18:38426745-38444055	3.60	0.70	5.00E-05	2.50E-03	5.13
Sox18	chr2:181404541-181406345	4.31	0.85	1.50E-04	6.36E-03	5.08
Csgalnact1	chr8:70880679-71259045	4.68	0.93	5.00E-05	2.50E-03	5.06
Adora2a	chr10:74779687-74797533	4.71	0.93	5.00E-05	2.50E-03	5.05
Mmrn2	chr14:35188689-35217472	12.11	2.41	5.00E-05	2.50E-03	5.01
Acer2	chr4:86520317-86566785	13.32	2.66	5.00E-05	2.50E-03	5.00
Thsd1	chr8:23337774-23371804	4.92	0.99	5.00E-05	2.50E-03	4.98
Tal1	chr4:114729365-114744360	31.18	6.31	5.00E-05	2.50E-03	4.94
Mpo	chr11:87607285-87617914	2.37	0.48	5.50E-04	1.78E-02	4.92
Afap1l1	chr18:61889053-61946316	17.92	3.69	5.00E-05	2.50E-03	4.85
Gngt2	chr11:95685813-95707045	21.37	4.46	1.75E-03	4.41E-02	4.79
Igf1	chr10:87321800-87399792	8.47	1.78	5.00E-05	2.50E-03	4.77

Table 3.17: Top 50 significantly upregulated genes in endothelial cells by fold change.

Gene	Locus	EC expression	ET expression	p-value	q-value	Fold Change
-	chr15:96991387-96991826	0.00	2.55	5.00E-05	2.50E-03	High
-	chr7:36255279-36256279	0.00	38.35	2.00E-03	4.91E-02	High
Gm10324	chr13:66214388-66223772	0.22	6.41	9.50E-04	2.74E-02	29.32
-	chr13:98252715-98274765	0.12	2.13	5.00E-05	2.50E-03	17.29
4930500J02Rik	chr2:104399333-104411586	0.10	1.60	9.50E-04	2.74E-02	16.68
-	chr13:98278330-98283216	0.11	1.78	6.50E-04	2.05E-02	16.53
Tdh	chr14:64111183-64127929	0.75	10.95	5.00E-05	2.50E-03	14.52
Gdf3	chr6:122555420-122560089	0.28	4.04	5.00E-04	1.67E-02	14.43
Gm2381	chr7:50067562-50122604	0.19	2.64	5.00E-04	1.67E-02	13.68
AU018091	chr7:3154659-3169204	0.88	11.07	5.00E-05	2.50E-03	12.56
Tdrd12	chr7:36278628-36322763	0.40	5.02	1.60E-03	4.12E-02	12.52
BC024139, Eppk1	chr15:75931917-75956986	0.12	1.47	2.50E-04	9.49E-03	12.30
Fgf4	chr7:152047290-152051148	0.36	4.35	5.00E-05	2.50E-03	12.21
-	chr9:118308486-118313594	0.15	1.77	2.50E-04	9.49E-03	11.78
Alox15	chr11:70157648-70165533	0.29	3.16	1.00E-04	4.56E-03	10.94
Utf1	chr7:147129754-147131011	0.85	9.21	2.00E-04	7.94E-03	10.78
Sptbn2	chr19:4711222-4752352	0.39	4.13	5.00E-05	2.50E-03	10.65
Gpa33	chr1:168060590-168096641	0.30	3.20	4.00E-04	1.40E-02	10.65
Folr1	chr7:109006844-109019302	0.43	4.59	5.00E-05	2.50E-03	10.62
Pou5f1	chr17:35642976-35647722	3.87	39.95	5.00E-05	2.50E-03	10.32
Esrp1	chr4:11259184-11313930	0.23	2.27	1.40E-03	3.74E-02	9.94
Slc28a1	chr7:88259684-88315302	0.23	2.23	2.00E-04	7.94E-03	9.80
Aire	chr10:77492766-77526360	0.31	3.04	5.00E-05	2.50E-03	9.79
Trap1a	chrX:135774764-135892277	1.38	13.49	5.00E-05	2.50E-03	9.79
Mcf2	chrX:57309132-57400820	0.16	1.51	1.55E-03	4.03E-02	9.72
Esrrb	chr12:87702066-87862578	0.65	6.28	5.00E-05	2.50E-03	9.71
Zfp42	chr8:44380420-44392363	1.04	10.10	5.00E-05	2.50E-03	9.70

Gm13242	chr4:145126547-145419626	0.55	5.21	1.60E-03	4.12E-02	9.56
Tfap2c	chr2:172375092-172384121	0.41	3.92	5.00E-05	2.50E-03	9.55
Sox1	chr8:12385770-12436732	0.34	3.21	5.00E-05	2.50E-03	9.43
Ano9	chr7:148287117-148303705	0.31	2.86	1.00E-04	4.56E-03	9.35
Wnt8b	chr19:44567961-44590041	0.30	2.83	1.00E-04	4.56E-03	9.33
Smc1b	chr15:84895118-84962387	0.23	2.12	1.00E-04	4.56E-03	9.28
Ap3b2	chr7:88605284-88638811	0.29	2.64	5.00E-05	2.50E-03	9.22
Rfx4	chr10:84218792-84369283	0.33	2.99	5.00E-05	2.50E-03	9.15
Nkx2-1	chr12:57632923-57637895	0.36	3.25	2.00E-04	7.94E-03	8.91
Gm13247	chr4:145651165-145696039	0.53	4.69	5.00E-05	2.50E-03	8.84
Fezf2	chr14:13174405-13179290	0.43	3.69	2.00E-04	7.94E-03	8.68
Wnt1	chr15:98620287-98624261	0.64	5.51	5.00E-05	2.50E-03	8.57
2410141K09Rik	chr13:66519049-66542054	2.10	18.01	1.15E-03	3.19E-02	8.57
Wnt7b	chr15:85365866-85424138	0.26	2.20	5.00E-05	2.50E-03	8.53
Dppa5a	chr9:78214860-78216006	35.20	298.69	5.00E-05	2.50E-03	8.49
Foxb1	chr9:69605516-69608747	0.19	1.63	1.70E-03	4.33E-02	8.38
Mlxipl	chr5:135582760-135614252	0.18	1.50	3.50E-04	1.26E-02	8.27
Nanog	chr6:122657585-122664639	2.20	18.08	5.00E-05	2.50E-03	8.23
Miat	chr5:112642247-112657968	1.81	14.87	5.00E-05	2.50E-03	8.20
Sim2	chr16:94085504-94348638	0.47	3.80	5.00E-05	2.50E-03	8.12
Pcdh8	chr14:80166578-80171119	0.64	5.20	5.00E-05	2.50E-03	8.11
Mpped1	chr15:83610452-83688904	0.21	1.72	5.00E-05	2.50E-03	8.07
D7Ertd143e	chr7:3217861-3221016	0.61	4.87	5.00E-05	2.50E-03	8.05

Associations of Differentially Regulated Genes with Endocardial Development

A number of the genes found to be over-expressed in the endocardium have known links to endocardial or cardiac development. The identification of these genes in our investigation supports the relevance of the embryoid body differentiation model of the endocardium to the *in vivo* differentiation process. This section discusses genes with known connections to endocardial or cardiac development, as well as *Oit3*, a gene not previously associated with heart development but exhibiting the maximal fold-change in the mRNA-seq data. In addition, the pluripotency factors that were unexpectedly found to be upregulated in the endothelium are discussed. The presence of pluripotency factors is unexpected because the FACS sorting protocol employed was originally expected to exclusively isolate differentiated endothelial cells.

Oit3, also known as *Lzp*, is the top hit from the mRNA-seq data by fold change and shows 17-fold upregulation in the endocardium compared to the endothelium. A literature search revealed no known function for *Oit3* in the heart or the endocardium. An *Oit3* null mouse has been generated [Yan et al., 2012] and has been found to survive to adulthood without any gross developmental abnormalities, suggesting the absence of major cardiac anomalies. Characterisation of the null mouse has revealed defects primarily in the kidney, although a full characterisation of the mouse is not reported in literature. Interestingly, earlier publications have demonstrated *Oit3* expression in heart muscle via qPCR but not via immunohistochemistry [Shen et al., 2009]. It is therefore possible that its endocardial specific expression detected in the RNA-seq data has previously been misattributed to the myocardium.

Ptprb, the fourth hit by fold-change, is a cell surface protein showing 10-fold up-regulation in the endocardium. *Ptprb* is also known as *Ve-ptp* [Wansleeben et al., 2011] and has been found to be essential for maintenance and remodelling, but not for formation, of blood vessels [Bäumer et al., 2006]. Furthermore, it has been shown to associate with *Tie2* [Bäumer et al., 2006], one of the two genes the knock-out of which completely ablates the endocardium, and negatively regulate its action [Winderlich et al., 2009] (Figure 3.29 B). *Ptprb* truncation has been shown to have widespread endothelial cell defects, but also specific endocardial defects, with absent trabeculation and failure of attachment of the endocardium to the myocardium, suggesting a more specific role in the heart [Bäumer et al., 2006] that has not been investigated. A specific role for *Ptprb* in the heart is further supported by its high expression in the OFT and developing heart valves [Dominguez

et al., 2007].

Cdh5, also known as *VE-Cadherin* and *CD144*, is a major component of endothelial adherens junctions [Dejana and Orsenigo, 2013]. *Cdh5* displays 8-fold up-regulation in the endocardium compared to the endothelium. Expression of *Cdh5* in the endocardium is known to occur [Narumiya et al., 2007], however, its over-expression in comparison to other endothelial cells has not been documented previously.

Sox7 is a transcription factor and member of the SoxF family of transcription factors that comprises *Sox7*, *Sox17* and *Sox18*. The SoxF family is known to have important roles in the development of the cardiovascular system [Francois et al., 2010] and *Sox7* in concert with *Sox18* have been found to be essential in cardiogenesis in *Xenopus laevis*. *Sox7* is upregulated 8-fold in the endocardium in the RNA-seq data, along with the other two members of the SoxF family (*Sox17* 2-fold, *Sox18* 2-fold). *Sox7* has previously been shown to be expressed in the developing endocardium at E8.75 [Sakamoto et al., 2007] among other mesodermaly derived endothelial tissues [Wat and Wat, 2014] and its deletion displays specific endocardial defects [Sakamoto et al., 2007].

Interestingly, *Sox7* has previously been shown to directly regulate expression of *Cdh5*, discussed above, by direct binding to its promoter region [Costa et al., 2012] and furthermore is directly regulated by *Etv2* [Behrens et al., 2014], which is also upregulated in the endocardium (3-fold). *Etv2*, also known as *Etsrp71*, is a member of the ETS family of transcription factors, a TF family independently identified in the context of this work as relevant to endocardial development (Section 3.3.9). This finding suggests that a pathway of *Etv2*, *Sox7*, *Cdh5* activation is active in the endocardium (Figure 3.29 A). *Etv2* has independently been shown to be indispensable in endocardial development and has been shown to be directly regulated by *Nkx2-5*, and in turn regulate *Tie2* [Ferdous et al., 2009] (Figure 3.29 A and B), one of the two tyrosine receptors that when knocked-out in mice results in complete and specific ablation of the endocardium.

Erg is a transcription factor and a member of the ETS family upregulated 8-fold in the endocardium. *Erg* has previously been shown to be a regulator of EMT in the valvular endocardium [Vijayaraj et al., 2012] and has been implicated in cardiac cushion formation [Schachterle et al., 2012]. Furthermore, *Erg* regulates expression of the *Gata4* transcription factor that is known to be expressed in several cardiovascular lineages (Figure 3.29 C). *Erg* physically interacts with *Klf2* and results in the up-regulation of *Flk1*, also known as *Kdr* and *Vegrf2*, (one of the markers of the MICP, see Section 1.3.3), which

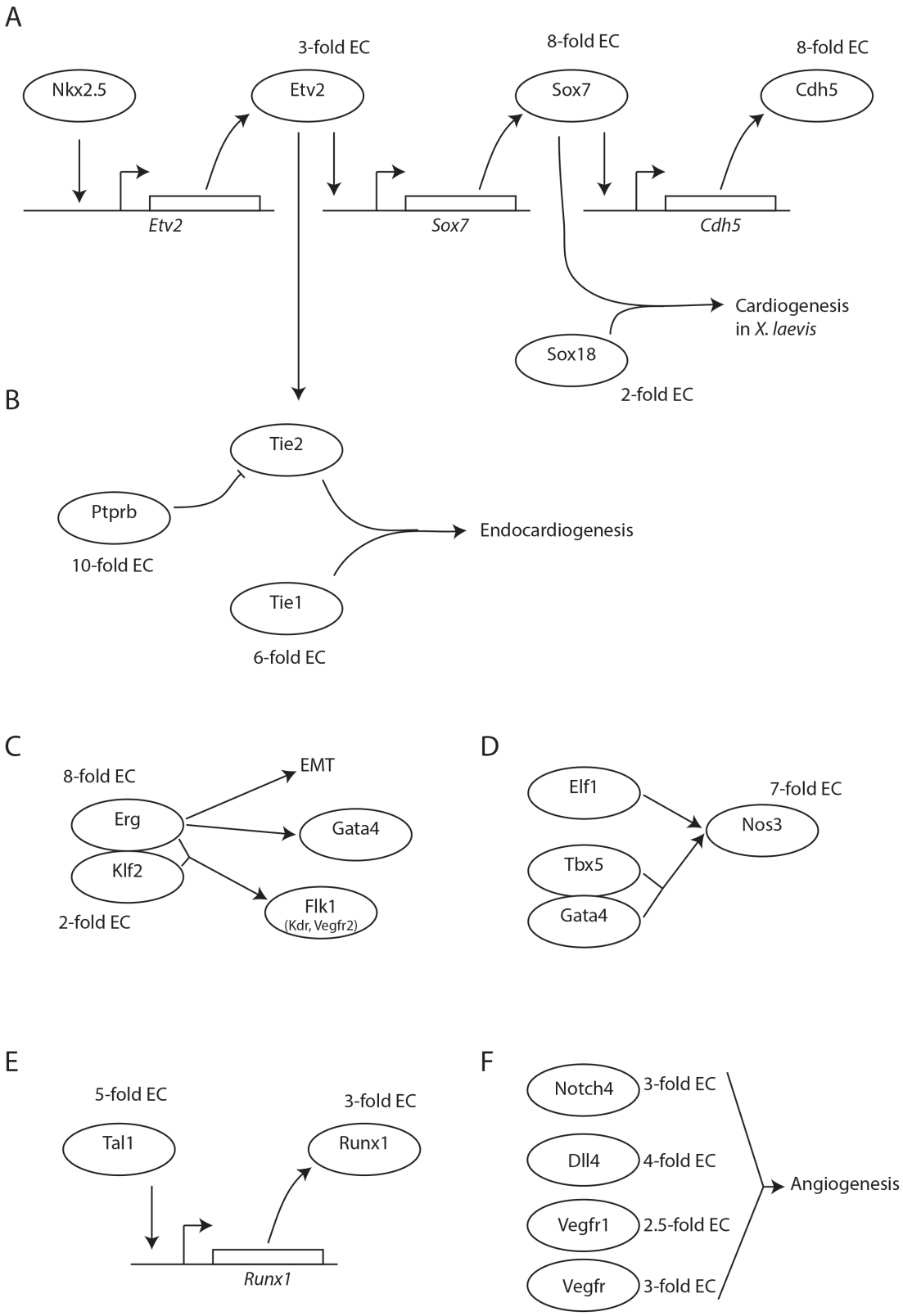


Figure 3.29: Summary of selected published gene relationships discussed in the main text. Genes upregulated in the mRNA-seq data are annotated by fold type and cell type of upregulation (EC endocardial, ET endothelial)

is also significantly upregulated in the endocardium 2-fold.

Nitric oxide synthetase 3 (*Nos3*), also known as endothelial Nos (*eNos*), is 7-fold upregulated in the endocardium (Figure 3.29 D). Nitric oxide (NO) is a signalling molecule implicated in a diverse repertoire of processes [Liu and Feng, 2012]. In contrast to nitric-oxide produced by *Nos1* and *Nos2*, nitric-oxide produced by *Nos3* has a role confined to intracellular signalling [Liu and Feng, 2012]. *Nos3* is known to have a role in heart development and a specific role in the endocardium. Mutations in *Nos3* result in congenital heart disease in the mouse and humans and *Nos3* null mice have been found to display atrial septal defects with high incidence [Nadeau et al., 2010] and defective bicuspid aortic valves [Liu and Feng, 2012]. In the context of the atrial septum formation, *Nos3* has been shown to have a specific role in endocardial cell survival and to be regulated by the genetically interacting *Tbx5* and *Gata4* transcription factors [Nadeau et al., 2010]. Interestingly, *Nos3* is known to be regulated by an ETS family member protein *Elf1* [Huang et al., 2005].

Tie1, an orphan tyrosine kinase receptor, shows 6-fold upregulation in the endocardium. *Tie1* has been strongly associated with endocardial development as it is one of the two genes (with *Tie2*) that when disrupted result in specific ablation of the endocardium [Puri et al., 1999]. Furthermore, *Tie1*-null mice display a reduction in endocardial cell numbers and heart developmental abnormalities [Dumont et al., 1994] (Figure 3.29 B).

Egfl7, also known as VE-statin, is upregulated 5-fold in the endocardium and is a known regulator of angiogenesis. *Egfl7* promotes angiogenesis via the *Stat3* signalling cascade [Chim et al., 2014]. *Stat3* is also significantly upregulated in the endocardium (1.5-fold), suggesting that the same pathway may be active in the endocardium. The gene body of *Eglf7* contains the gene for *microRNA-127*. *MicroRNA-127*, is directly regulated by ETS family members *Ets1* and *Ets2* [Harris et al., 2010] that are both upregulated in the endocardial mRNA-seq data and has been implicated in vascular development [Ásgeirsdóttir et al., 2012].

Tal1, a DNA binding transcription factor [Deleuze et al., 2007] that is upregulated 5-fold in the endocardium, has also been previously linked to endocardial development. A *Tal1 Danio rerio* mutant with abnormal segregation of the endocardium has been described [Bussmann et al., 2007] and furthermore *Tal1* has been found to influence the distribution of endocardial cell junctions *in vivo* and also be required for the maintenance

of the cell identity for the endocardium in the mouse [Schumacher et al., 2013]. In the context of the hemangioblast, the progenitor of haematopoietic and epithelial cells that precedes formation of the endocardium (Section 1.3.1), *Tal1* has been shown to directly target the promoter region of *Runx1* [Landry et al., 2008] (Figure 3.29 E), a TF that shows 3-fold upregulation in the endocardium. Ablation of *Tal1* has been found to result in *Runx1* downregulation [Van Handel et al., 2012].

In addition to the aforementioned factors, all four factors known to direct angiogenesis are upregulated in the endocardium. These factors comprise *Notch4* (upregulated 3-fold), *Dll4* (upregulated 4-fold), *Vegfr1* (also known as *Flt1*, upregulated 2.5-fold) and *Vegfr* (also known as *Kdr* and *Flk*, upregulated 3-fold) (Figure 3.29 F). These four factors are involved in angiogenesis by directing the growth of developing vessels [Jakobsson et al., 2010] and computational models suggest that they are sufficient for the patterning of vessels, in response to exogenous gradients. Establishment of signalling via this angiogenic pathway is by ETS transcription factors [Wythe et al., 2013]. Specifically the activity of the *Dll4* enhancer is regulated by ETS factors *Etv2* and *Erg*, both of which are upregulated in the endocardium in the mRNA-seq data.

Finally, it must be noted that NFATc1 – the marker of the endocardium used for FAC sorting – is upregulated 4-fold in the endocardium.

Several genes are upregulated in the endothelium, and conversely downregulated in the endocardium. Given that the endothelium examined represents the average background population of endothelial cells in the embryoid bodies and no particular subpopulation of the endothelium, genes that appear to be upregulated in this tissue could be upregulated in any subcomponent of it, or be specifically downregulated in the endocardium. It is therefore difficult to draw specific conclusions from the group of candidate genes upregulated in the endothelium. It is however interesting to note that two of the four known pluripotency factors appear highly differentially expressed in the endothelium compared to the endocardium.

Specifically, *Pou5f1* (also known as *Oct3/4*) is 8-fold upregulated in the endothelium and *Sox2* is 5-fold upregulated. Furthermore, *Nanog*, [Chambers and Tomlinson, 2009] is 8-fold upregulated. The upregulated pluripotency factors are all part of the LIF pathway [Niwa et al., 2009] that is responsible for maintenance of pluripotency of ES cells during culture on MEFs (Section 2.1). Expression of these genes, suggests that MEFs may survive into the hanging drop EB culture, despite their irradiation and depletion. Selection for

$CD31^+$ that might be expected to result in the depletion of pluripotent cells is probably not completely effective as a result of CD31 expression by pluripotent cells [Robson et al., 2001].

3.3.6 Differential Promoter Usage and Alternative Splicing

Although the aim of this investigation did not explicitly include the identification of differentially spliced transcripts and the experimental design was not such as to allow complete identification of these events, this analysis was possible and was therefore performed.

A limited number of transcripts in the dataset were found to display differential promoter usage or alternative splicing. These are displayed in Table 3.18 along with locus position, uncorrected and corrected p-values.

Four genes have been found to show alternative splicing. *Afap1l1* (Actin filament associated protein 1-like 1), a gene encoding for a protein of unknown function, has recently been described as a regulator of cellular morphology in colorectal cancer [Takahashi et al., 2014]. Different Afap1l1 isoforms may potentially be responsible for subtle cytoskeletal differences between the endocardium and the endothelium. Myocd (Myocardin) is described as essential for myocardial survival and is known to be a transcriptional regulator in smooth muscle cells [Huang et al., 2009]. However, a role in the pathogenesis of vascular disease has been proposed [Zheng, 2014] and its role in the endothelium is corroborated by our data. Lphn2 is a putative uncharacterised protein and no information of Ccser2 exists in literature.

Three genes exhibit alternative promoter usage. *Sall2* (sal-like 2) is described as a transcription factor in literature and has been implicated in cancer [Farkas et al., 2013], but is not otherwise known to have a role in heart or vascular development. *Egfl7* (epidermal growth factor-like protein 7) is a secreted angiogenic factor with well characterised role in vascular development and a known role in endothelial proliferation [Nichol and Stuhlmann, 2011]. Although *Egfl7* is known to make use of alternative promoter sites, an endocardial-specific isoform of the protein product has not been previously described and its identification in this assay warrants further investigation. Gpm6b (neuronal membrane glycoprotein M6-b) has been previously been shown to display sexually dimorphic expression in the heart and may therefore represent a false positive [Isensee et al., 2008].

Of the above hits, *Afap1l1*, *Sall2*, *Eglf6* and *Gpm6b* also appear to have a difference in expression levels between the two cell types in the differential expression analysis.

Table 3.18: All transcripts showing alternative promoter usage or alternative splicing between endocardial and endothelial cells.

	Gene	Locus	p-value	q-value
Splicing				
	Afap1l1	chr18:61889053-61946316	5.00E-05	0.0234125
	Lphn2	chr3:148478549-148617605	5.00E-05	0.0234125
	Myocd	chr11:64990071-65083491	5.00E-05	0.0234125
	Ccser2	chr14:37688121-37781950	5.00E-05	0.0234125
Promoter				
	Sall2	chr14:52930851-52948345	5.00E-05	0.02115
	Egfl7	chr2:26436575-26448202	5.00E-05	0.02115
	Gpm6b	chrX:162676874-162826965	0.00015	0.0423

3.3.7 Gene Ontology Term Overrepresentation Analysis

The set of differentially expressed genes was assessed for overrepresentation of Gene Ontology (GO) terms using a pruned tree approach with GO-Elite [Zambon et al., 2012] as outlined in Section 2.7.7. The ten most significant terms of each analysis split by Biological Process, Cellular Component and Molecular Function can be found in Tables 3.19, 3.20 and 3.21. Full results can be found in Tables A.5, A.6 and A.7 of the Appendix.

The results are consistent with the expectations from these cell lines with terms such as "developmental process", "cell adhesion", "regulation of developmental process", "regulation of cell migration", "regulation of response to stimulus" comprising the most significantly overrepresented terms. In particular, the prominence of the term "developmental process" is notable as it is the most significant term in the biological process GO term analysis with a p-value of 0.01 and reaffirms with the active developmental process endocardial cells are undergoing.

It is of interest to note that the term "response to fluid shear stress" appears in the list of significantly overrepresented GO terms suggesting that the haemodynamic environment may have an influence on the transcriptome of the endocardium. Genes associated with this term comprise *Cited2*, *Ets1*, *Mef2c*, *Nos3*, *Smad6* and *Tgfb1*. This would not be unprecedented [Banjo et al., 2013] and would support a model where the endocardium arises from the the endothelium and the differences from other endothelium are primarily the result of the local haemodynamic environment.

The differentially expressed genes were stratified by upregulation in either endocardial or endothelial cells and GO term analysis was repeated in both sets separately. The most significant results of the analysis for genes upregulated in the endocardium and

endothelium are shown in Tables 3.22 through 3.27. Complete tables of the results of the analysis can be found in Tables A.8 through to A.10 and A.11 through to A.13 of the Appendix.

It is of interest to note that whereas the genes upregulated in the endothelium are associated with general developmental terms (such as "anatomical structure development" and "embryonic pattern specification"), endocardially upregulated genes appear to have more specific enrichment for haematopoietic development ("regulation of erythrocyte development", "regulation of mast cell differentiation") while still showing overrepresentation of endothelial genes ("endothelial cell differentiation", "angiogenesis"). The stratified analysis, confirms that the term "response to fluid shear stress" is specifically associated with endocardially upregulated genes.

Table 3.19: Ten most significantly overrepresented, pruned, biological process GO terms from the differential expression analysis. Terms are sorted by ascending adjusted p-value and then by descending Z-score.

Ontology Name (Ontology-ID)	Percent Changed	Z-Score	Adjusted P-value
developmental process(GO:0032502)	8.06	16.86	1.01E-02
cell adhesion(GO:0007155)	13.56	14.43	1.01E-02
regulation of developmental process(GO:0050793)	9.63	12.97	1.01E-02
regulation of cell migration(GO:0030334)	15.00	12.17	1.01E-02
regulation of response to stimulus(GO:0048583)	7.94	11.07	1.01E-02
positive regulation of biological process(GO:0048518)	6.71	11.05	1.01E-02
regulation of cell proliferation(GO:0042127)	9.29	10.83	1.01E-02
response to external stimulus(GO:0009605)	9.87	10.55	1.01E-02
negative regulation of cellular process(GO:0048523)	6.88	10.17	1.01E-02
negative regulation of locomotion(GO:0040013)	20.35	10.17	1.01E-02

Table 3.20: Ten most significantly overrepresented, pruned, cellular component GO terms from the differential expression analysis. Terms are sorted by ascending adjusted p-value and then by descending Z-score.

Ontology Name(Ontology-ID)	Percent Changed	Z-Score	Adjusted P-value
extracellular region part(GO:0044421)	9.31	11.34	1.01E-02
plasma membrane part(GO:0044459)	7.58	11.03	1.01E-02
cell surface(GO:0009986)	13.41	10.80	1.01E-02
plasma membrane(GO:0005886)	6.61	10.59	1.01E-02
neuron projection(GO:0043005)	9.38	8.04	1.01E-02
dendrite terminus(GO:0044292)	75.00	8.02	1.01E-02
extracellular region(GO:0005576)	6.71	7.74	1.01E-02
membrane raft(GO:0045121)	10.14	5.54	1.01E-02
cell periphery(GO:0071944)	21.74	4.95	1.01E-02
extrinsic to membrane(GO:0019898)	10.92	4.67	1.01E-02

Table 3.21: Ten most significantly overrepresented, pruned, molecular function GO terms from the differential expression analysis. Terms are sorted by ascending adjusted p-value and then by descending Z-score.

Ontology Name(Ontology-ID)	Percent Changed	Z-Score	Adjusted P-value
protein binding(GO:0005515)	4.93	11.59	1.01E-02
pattern binding(GO:0001871)	18.07	10.69	1.01E-02
calcium ion binding(GO:0005509)	10.87	10.52	1.01E-02
extracellular matrix binding(GO:0050840)	28.95	8.85	1.01E-02
RNA polymerase II regulatory region sequence-specific DNA binding(GO:0000977)	22.22	8.41	1.01E-02
Wnt receptor activity(GO:0042813)	35.00	7.93	1.01E-02
chemorepellent activity(GO:0045499)	60.00	7.09	1.01E-02
retinoic acid binding(GO:0001972)	44.44	6.91	1.01E-02
transmembrane receptor protein kinase activity(GO:0019199)	16.67	6.87	1.01E-02
icosanoid binding(GO:0050542)	50.00	6.40	1.01E-02

Table 3.22: Ten most significantly overrepresented, pruned, biological process GO terms from the genes upregulated in the endocardium. Terms are sorted by ascending adjusted p-value and then by descending Z-score.

Ontology Name (Ontology-ID)	Percent Changed	Z-Score	Adjusted P-value
regulation of erythrocyte differentiation(GO:0045646)	32.26	14.55	1.20E-02
regulation of mast cell differentiation(GO:0060375)	100.00	14.46	1.20E-02
endothelial cell differentiation(GO:0045446)	55.56	13.75	1.20E-02
regulation of gamma-delta T cell activation(GO:0046643)	66.67	13.53	1.20E-02
angiogenesis(GO:0001525)	12.90	13.33	1.20E-02
JAK-STAT cascade involved in growth hormone signaling pathway(GO:0060397)	75.00	12.46	1.20E-02
regulation of cell motility(GO:2000145)	9.07	12.28	1.20E-02
response to fluid shear stress(GO:0034405)	37.50	12.22	1.20E-02
regulation of angiogenesis(GO:0045765)	13.64	11.93	1.20E-02
hemopoiesis(GO:0030097)	15.79	11.89	1.20E-02

Table 3.23: Ten most significantly overrepresented, pruned, cellular component GO terms from the genes upregulated in the endocardium. Terms are sorted by ascending adjusted p-value and then by descending Z-score.

Ontology Name (Ontology-ID)	Percent Changed	Z-Score	Adjusted P-value
plasma membrane(GO:0005886)	3.25	8.90	1.20E-02
cell periphery(GO:0071944)	21.74	8.26	1.20E-02
cell surface(GO:0009986)	6.15	7.65	1.20E-02
plasma membrane part(GO:0044459)	3.24	7.12	1.20E-02
acrosomal membrane(GO:0002080)	25.00	6.92	1.20E-02
transport vesicle(GO:0030133)	10.71	5.90	1.20E-02
extracellular region part(GO:0044421)	3.26	5.26	1.20E-02
cell projection(GO:0042995)	3.21	5.01	1.20E-02
cell fraction(GO:0000267)	2.66	3.70	1.20E-02
intrinsic to Golgi membrane(GO:0031228)	11.11	4.93	2.18E-02

Table 3.24: Ten most significantly overrepresented, pruned, molecular function GO terms from the genes upregulated in the endocardium. Terms are sorted by ascending adjusted p-value and then by descending Z-score.

Ontology Name (Ontology-ID)	Percent Changed	Z-Score	Adjusted P-value
RNA polymerase II regulatory region sequence-specific DNA binding(GO:0000977)	15.87	9.73	1.20E-02
protein binding(GO:0005515)	2.32	9.73	1.20E-02
pattern binding(GO:0001871)	7.23	6.37	1.20E-02
calcium ion binding(GO:0005509)	4.35	6.17	1.20E-02
core promoter sequence-specific DNA binding(GO:0001046)	15.38	6.03	1.20E-02
sequence-specific DNA binding transcription factor activity(GO:0003700)	3.54	5.26	1.20E-02
guanyl-nucleotide exchange factor activity(GO:0005085)	6.33	5.25	1.20E-02
enzyme activator activity(GO:0008047)	4.49	4.80	1.20E-02
chromatin binding(GO:0003682)	4.38	4.01	1.20E-02
receptor signaling protein activity(GO:0005057)	6.41	3.74	1.20E-02

Table 3.25: Ten most significantly overrepresented, pruned, biological process GO terms from the genes upregulated in the endothelium. Terms are sorted by ascending adjusted p-value and then by descending Z-score.

Ontology Name(Ontology-ID)	Percent Changed	Z-Score	Adjusted P-value
cell-cell adhesion(GO:0016337)	14.44	15.26	2.07E-02
anatomical structure development(GO:0048856)	5.84	13.59	2.07E-02
embryonic pattern specification(GO:0009880)	25.49	12.40	2.07E-02
forebrain anterior/posterior pattern formation(GO:0021797)	66.67	11.66	2.07E-02
anatomical structure morphogenesis(GO:0009653)	6.52	11.65	2.07E-02
axon guidance(GO:0007411)	15.75	11.51	2.07E-02
axis specification(GO:0009798)	20.00	11.15	2.07E-02
multicellular organismal development(GO:0007275)	6.83	11.02	2.07E-02
tube formation(GO:0035148)	16.16	10.46	2.07E-02
cellular developmental process(GO:0048869)	4.87	9.78	2.07E-02

Table 3.26: Ten most significantly overrepresented, pruned, cellular component GO terms from the genes upregulated in the endothelium. Terms are sorted by ascending adjusted p-value and then by descending Z-score.

Ontology Name(Ontology-ID)	Percent Changed	Z-Score	Adjusted P-value
extracellular matrix(GO:0031012)	9.94	10.87	2.07E-02
cell-cell junction(GO:0005911)	9.06	8.63	2.07E-02
extracellular region(GO:0005576)	4.67	8.30	2.07E-02
fibrillar collagen(GO:0005583)	33.33	8.00	2.07E-02
extracellular space(GO:0005615)	5.60	7.80	2.07E-02
cell surface(GO:0009986)	7.26	7.54	2.07E-02
lateral plasma membrane(GO:0016328)	20.69	7.44	2.07E-02
axon(GO:0030424)	8.37	6.82	2.07E-02
apical plasma membrane(GO:0016324)	7.73	6.40	2.07E-02
plasma membrane(GO:0005886)	3.36	6.18	2.07E-02

Table 3.27: Ten most significantly overrepresented, pruned, molecular function GO terms from the genes upregulated in the endothelium. Terms are sorted by ascending adjusted p-value and then by descending Z-score.

Ontology Name(Ontology-ID)	Percent Changed	Z-Score	Adjusted P-value
chemorepellent activity(GO:0045499)	60.00	9.55	2.07E-02
extracellular matrix binding(GO:0050840)	21.05	8.69	2.07E-02
transmembrane-ephrin receptor activity(GO:0005005)	50.00	8.66	2.07E-02
pattern binding(GO:0001871)	10.84	8.51	2.07E-02
calcium ion binding(GO:0005509)	6.52	8.46	2.07E-02
inorganic anion exchanger activity(GO:0005452)	33.33	8.00	2.07E-02
Wnt receptor activity(GO:0042813)	25.00	7.60	2.07E-02
axon guidance receptor activity(GO:0008046)	37.50	7.40	2.07E-02
heparan sulfate proteoglycan binding(GO:0043395)	28.57	7.34	2.07E-02
Wnt-protein binding(GO:0017147)	20.00	6.66	2.07E-02

Table 3.28: Transcription Factors upregulated in endocardial cells.

Asb4	Foxh1	Myb	Shank3
Bcl6b	Gata1	Myc	Smad6
Cbfa2t3	Gata2	Nfatc1	Sox17
Cited2	Gfi1b	Nfe2	Sox18
Elk3	Hhex	Notch4	Sox7
Erg	Hoxb3	Ppp1r13b	Stat3
Ets1	Ikzf1	Ppp1r16b	Stat5a
Ets2	Lmo2	Rab11a	Stat5b
Etv2	Lyl1	Rreb1	Tal1
Fli1	Mef2c	Runx1	Zfp711
			Zfpm1

3.3.8 Identification of Differentially Regulated Transcription Factors

In order to derive a list of differentially expressed transcription factors, the set of differentially expressed genes was searched for genes that are known to act as transcription factors. The intersection of the differentially expressed genes with the union of three independent annotated lists of transcription factors was obtained.

Differentially regulated genes between the endocardium and the endothelium that were annotated with the GO Term "sequence-specific DNA binding transcription factor activity" (GO:0003700) were obtained. This term was selected as the most relevant term to TFs after manual inspection of the GO ontology. A custom MySQL (see Appendix Section C.4) query was performed against a local copy of the GO database.

The list of transcription factors identified above was further expanded by obtaining the intersection of the differentially regulated genes with the list non-redundant transcription factors identified by Kanamori and colleagues [Kanamori et al., 2004] and independently with the list of mouse only transcription factors generated by Zhang and colleagues [Zhang et al., 2012], as described in Section 2.7.8.

Upregulated and downregulated transcription factors are presented in Tables 3.28 and 3.29 respectively. These lists can be used for prioritisation of genes more likely to have a regulatory as opposed to effector role.

Table 3.29: Transcription Factors downregulated in endocardial cells.

Aire	Gm13242	Pax6	Sox3
Arnt2	Grhl2	Plagl1	Tcea3
Bnc1	Lhx2	Pou3f1	Tfap2a
Dbx1	Lmx1a	Pou5f1	Tfap2c
Dmrt1	Mlxipl	Rfx2	Tox3
Elavl2	Mycl1	Rfx4	Utf1
Elf3	Nanog	Sall1	Ybx2
Esrrb	Nkx2-1	Sall2	Zbtb16
Fezf2	Nr2f2	Sim2	Zfhx4
Foxa2	Nr5a2	Six3	Zfp296
Foxb1	Otx1	Sox1	Zfp42
Gli1	Otx2	Sox2	Zfp534
Gm13051	Pax3	Sox21	Zic2
			Zic3

3.3.9 Transcription Start Site Motif Analysis

The identification of a single or small subset of transcription factors (TF) or sequence elements that are involved in the restricted phenotypic differences of endocardial cells would be of particular interest because it could constitute a starting point for the identification of upstream regulators and would provide insight into the processes that functionally differentiate the endocardium from other endothelium. To this end, genomic regions surrounding the TSS of the genes upregulated in the endocardium were examined further. Some of the regulators of endocardial identity are likely to regulate transcription by direct binding to specific sequences of the TSS in downstream genes. On this basis, the sequence of the TSSs of differentially regulated genes are likely to be enriched in specific sequences bound by these factors.

In order to identify these potentially overrepresented sequences, motif analysis was performed on the genomic regions adjacent to the TSS of genes upregulated in the endocardium. A search interval for the motif analysis was defined on the basis of the binding pattern of TFs as determined by re-analysis of the ENCODE ChIP-seq data, described in Section 2.7.9. Briefly, ENCODE peak data [The ENCODE Project Consortium, 2012] for twelve TFs were downloaded from the UCSC browser and the distribution of the peaks in relation to the nearest TSS, within 20 kb, was plotted (See Figure 3.30 and A.1 through to A.5). Visual examination of the plots revealed that although TFs display varying patterns of binding, binding frequency around TSSs is close to background levels no more than 5 kb from either side of the TSS for the TFs examined.

A 10 kb window surrounding the TSS of upregulated genes was therefore used for motif analysis using DREME [Bailey et al., 2009]. This window was further subdivided into 1 kb intervals overlapping by 500 bps. Overlapping intervals were employed so as to ensure that motifs present at the boundary of two intervals can be identified by the analysis. Motif analysis was performed on these intervals against a background of matched sequences obtained from TSSs of stably expressed genes, as described in Section 2.7.10 with a stringency cut-off of E-value less than 0.05. The analysis identified 22 unique motifs in the TSSs of upregulated genes.

The TOMTOM tool was subsequently used to match overrepresented motifs identified above to known motifs of TFs. The TOMTOM tool reports all known motifs matching each supplied motif with different degrees of similarity and confidence in the match. A high level of stringency was required to report a match as significant in the context of

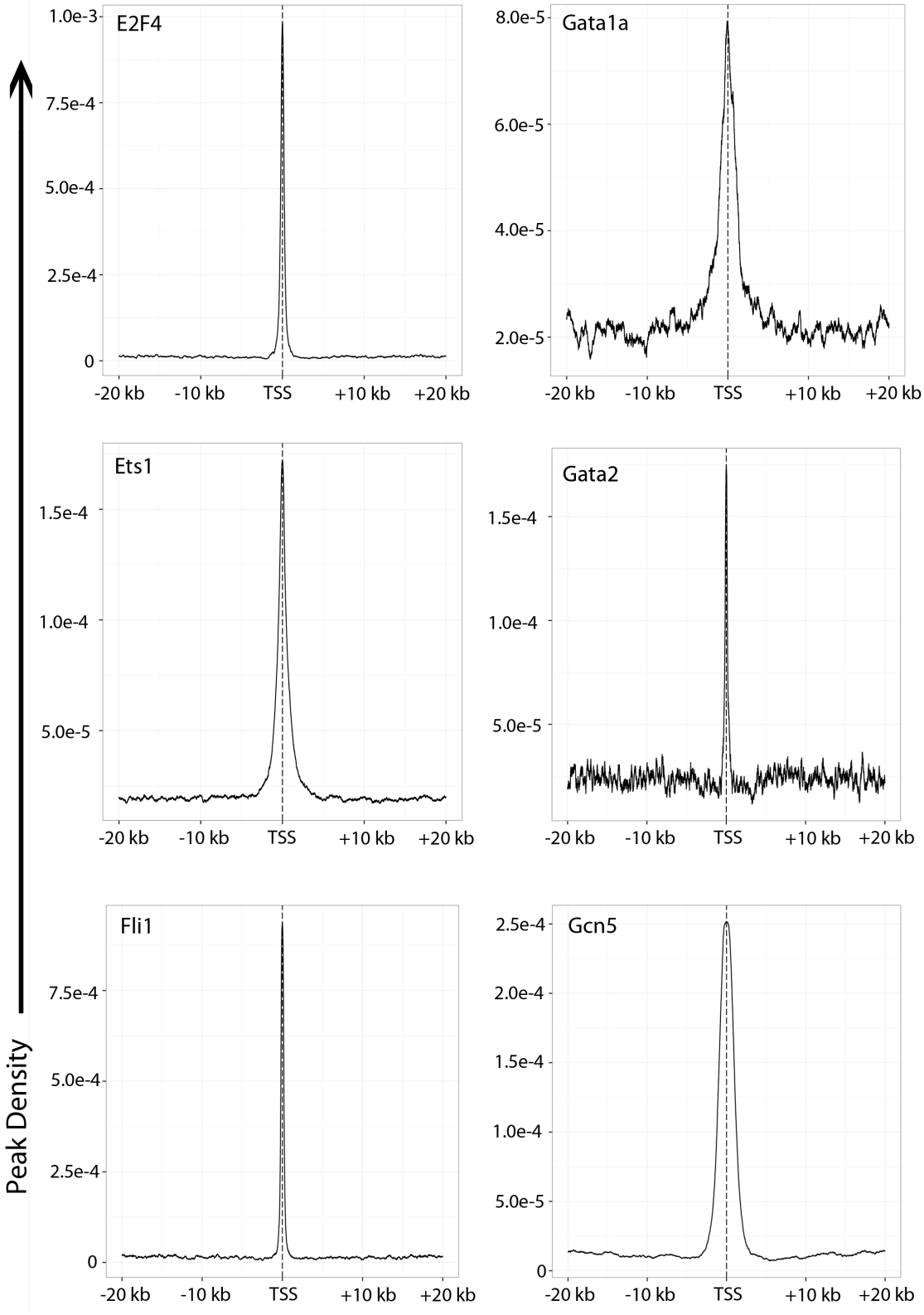


Figure 3.30: Binding distribution of six representative transcription factors (E2F4, Gata1a, Ets1, Gata2, Fli1, Gcn5) prepared using publicly ENCODE data. The vertical axis denotes peak density - the fraction of all the peaks at each position. The horizontal axis denotes distance from the nearest TSS.

this analysis; specifically, a motif was reported only if its E-value and q-value were both smaller than 0.05. The TOMTOM analysis identified 45 known motifs that match the 22 unique motifs identified by the DREME analysis, with some newly identified motifs matching multiple known ones. The full results of this analysis are presented in Table 3.30.

Some of the identified motifs correspond to factors previously associated with cardiovascular development, supporting the relevance of this analysis in the discovery of factors that regulate endocardial development. Specifically, the motifs of five members (*Ehf*, *Erg*, *Ets1*, *Fli1* and *Elf1*) of the ETS family of transcription factors were identified as matching overrepresented motifs (CTTCCKS, CTTCCTS, ASAGGAAR, RCAGGAA, GGAAGGW). The ETS family of TFs has been previously associated with cardiac and coronary development [Vijayaraj et al., 2012], regulation of the master cardiac enhancer *Gata4* [Schachterle et al., 2012] and recently heart mesoderm specification [Nie and Bronner, 2015].

Further to the ETS family motifs, the motifs of *Zic3*, *Mzf1* and *Brachyury* were identified. *Zic3* mutations are known to result in congenital heart disease [Cowan et al., 2014], whereas *Brachyury* (*T*) is a known regulator of mesoderm specification and *Mzf1* has been found to modulate mouse cardiogenesis along with *Nkx2.5* [Doppler et al., 2014].

In addition to the above factors, several other factors were identified and are shown in Table 3.30 that may also be involved in endocardial development. The most prominent of which was *Ehf* that was identified with the minimum q-value of 3.3e-5. *Ehf* has not previously been associated with endocardial development.

In order to further confirm the biological relevance of the motifs identified and to validate the discovery analysis, the distribution of the identified motifs in the promoters of both upregulated and stably expressed genes was examined. All the motifs identified were more frequently present in the sequences of the upregulated promoters compared to stably expressed promoters (Table 3.31). Furthermore, the distribution of 12 of the 22 motifs (CCCCAYCC, AGGGGVC, GGAAGGW, CCCMYTCC, CACGBGC, CCCTGRR, CTGRGGC, CTCTSMCC, BTCCCCA, CTTCCKC, CAGGGAS, AGGAGGD) displayed a distribution significantly different (Kolmogorov-Smirnov Test) from a uniform distribution suggesting that the positioning of the motifs is significant and that the motifs are likely to be of functional importance. The results of this analysis are shown in Figures 3.31, 3.32 and Tables 3.31 and 3.34.

Table 3.30: Overrepresented motifs identified by DREME in the promoter regions of genes overexpressed in the endocardium that have been identified by TOMTOM as matching to a known motif sequence.

Window	Motif	Motif Name	Motif Alt Name	DREME E-value	TOMTOM p-value	TOMTOM E-value	TOMTOM q-value
-5000	CACGBGC	MA0281.1	CBF1	9.60E-04	9.18E-06	8.99E-03	3.71E-03
-5000	CACGBGC	MA0569.1	MYC4	9.60E-04	9.18E-06	8.99E-03	3.71E-03
-5000	CACGBGC	MA0449.1	h	9.60E-04	1.07E-05	1.05E-02	3.71E-03
-5000	CACGBGC	MA0566.1	MYC2	9.60E-04	1.17E-05	1.15E-02	3.71E-03
-5000	CACGBGC	MA0357.1	PHO4	9.60E-04	3.63E-05	3.55E-02	7.72E-03
-5000	CACGBGC	MA0464.1	Bhlhe40	9.60E-04	3.65E-05	3.58E-02	7.72E-03
-5000	CACGBGC	MA0568.1	MYC3	9.60E-04	4.81E-05	4.71E-02	7.86E-03
-5000	GGTGTSA	MA0009.1	T	1.50E-02	8.54E-06	8.36E-03	1.67E-02
-5000	CTGTGS	MA0332.1	MET28	4.70E-02	1.15E-05	1.13E-02	2.25E-02
-5000	CTGTGS	MA0335.1	MET4	4.70E-02	4.66E-05	4.56E-02	4.56E-02
-4500	CCCMYTCC	MA0516.1	SP2	1.50E-06	2.82E-05	2.76E-02	4.50E-02
-4500	GTGACCM	MA0016.1	usp	3.00E-03	1.90E-05	1.86E-02	3.70E-02
-4500	GTGACCM	MA0016.1	usp	3.00E-03	2.02E-05	1.98E-02	3.94E-02
-3500	CCCCAYCC	MA0599.1	KLF5	1.70E-03	1.23E-05	1.20E-02	1.57E-02
-3500	CCCCAYCC	MA0039.2	Klf4	1.70E-03	1.77E-05	1.73E-02	1.57E-02
-3500	CCCCAYCC	UP00099_2	Ascl2_secondary	1.70E-03	2.44E-05	2.39E-02	1.57E-02
-3500	CCCCAYCC	MA0599.1	KLF5	1.70E-03	1.97E-05	1.93E-02	2.47E-02
-3500	CCCCAYCC	MA0039.2	Klf4	1.70E-03	3.22E-05	3.15E-02	2.47E-02
-3500	CCCCAYCC	UP00099_2	Ascl2_secondary	1.70E-03	3.81E-05	3.73E-02	2.47E-02
-3000	GGAAGGW	MA0149.1	EWSR1-FLI1	3.50E-02	1.02E-05	9.95E-03	1.99E-02
-2500	AGGGGVC	MA0366.1	RGM1	2.80E-05	2.12E-05	2.07E-02	4.13E-02
-2500	AGGGGVC	MA0342.1	MSN4	2.80E-05	4.91E-05	4.80E-02	4.79E-02
-2500	CTTCCKC	MA0080.3	Spi1	2.50E-04	4.32E-05	4.23E-02	4.77E-02
-2500	CTTCCKC	MA0598.1	EHF	2.50E-04	4.89E-05	4.78E-02	4.77E-02
-2000	CTTCCTS	MA0598.1	EHF	1.10E-02	1.70E-08	1.66E-05	3.30E-05

-2000	CTTCCTS	MA0474.1	Erg	1.10E-02	3.34E-06	3.27E-03	3.25E-03
-2000	CTTCCTS	MA0098.2	Ets1	1.10E-02	2.26E-05	2.21E-02	1.18E-02
-2000	CTTCCTS	MA0475.1	FLI1	1.10E-02	2.42E-05	2.37E-02	1.18E-02
-2000	CTTCCTS	MA0473.1	ELF1	1.10E-02	3.39E-05	3.32E-02	1.32E-02
-1500	CCCTGRR	MA0524.1	TFAP2C	2.20E-10	2.31E-05	2.26E-02	4.49E-02
-1500	CCCTGRR	MA0154.2	EBF1	2.20E-10	4.68E-05	4.58E-02	4.55E-02
-1000	ASAGGAAR	MA0081.1	SPIB	6.80E-11	1.51E-05	1.48E-02	1.59E-02
-1000	ASAGGAAR	MA0474.1	Erg	6.80E-11	1.63E-05	1.59E-02	1.59E-02
-1000	ASAGGAAR	MA0080.3	Spi1	6.80E-11	4.29E-05	4.20E-02	2.79E-02
-1000	BACCCCA	MA0595.1	SREBF1	3.80E-08	1.79E-05	1.75E-02	1.74E-02
-1000	BACCCCA	MA0596.1	SREBF2	3.80E-08	1.79E-05	1.75E-02	1.74E-02
-1000	BTCCCCA	MA0056.1	MZF1_1-4	3.80E-08	1.74E-05	1.70E-02	3.40E-02
-500	RCAGGAA	MA0475.1	FLI1	2.20E-09	4.08E-06	3.99E-03	7.37E-03
-500	RCAGGAA	MA0474.1	Erg	2.20E-09	7.56E-06	7.40E-03	7.37E-03
-500	RCAGGAA	MA0098.2	Ets1	2.20E-09	1.36E-05	1.33E-02	8.85E-03
-500	HTGGGGA	MA0056.1	MZF1_1-4	8.20E-06	2.54E-05	2.48E-02	4.97E-02
-500	CACCCTK	MA0112.2	ESR1	2.00E-05	2.49E-05	2.44E-02	4.88E-02
-500	RCAGGAA	MA0475.1	FLI1	2.20E-09	4.08E-06	3.99E-03	7.37E-03
-500	RCAGGAA	MA0474.1	Erg	2.20E-09	7.56E-06	7.40E-03	7.37E-03
-500	RCAGGAA	MA0098.2	Ets1	2.20E-09	1.36E-05	1.33E-02	8.85E-03
-500	HTGGGGA	MA0056.1	MZF1_1-4	8.20E-06	2.54E-05	2.48E-02	4.97E-02
-500	CACCCTK	MA0112.2	ESR1	2.00E-05	2.49E-05	2.44E-02	4.88E-02
0	GGGTGTS	MA0270.1	AFT2	1.50E-02	1.33E-06	1.30E-03	2.60E-03
0	GGGTGTS	MA0493.1	Klf1	1.50E-02	2.20E-05	2.15E-02	2.15E-02
500	CAGGGAS	MA0540.1	DPY-27	2.10E-07	1.64E-05	1.60E-02	3.21E-02
500	CTGRGGC	MA0524.1	TFAP2C	5.10E-05	8.86E-06	8.68E-03	1.72E-02
500	CTGRGGC	MA0003.2	TFAP2A	5.10E-05	2.05E-05	2.00E-02	1.99E-02
500	CACAGMAG	UP00102_2	Zic1_secondary	2.10E-02	2.16E-05	2.11E-02	4.22E-02
500	CACAGMAG	UP00006_2	Zic3_secondary	2.10E-02	4.49E-05	4.40E-02	4.40E-02
2000	AGGGGS	MA0342.1	MSN4	7.60E-08	4.89E-05	4.79E-02	4.72E-02
2000	AGGGGS	MA0366.1	RGM1	7.60E-08	4.89E-05	4.79E-02	4.72E-02

2000	BCGGRG	MA0348.1	OAF1	4.50E-05	2.04E-05	2.00E-02	3.93E-02
2000	CCMCAC	MA0493.1	Klf1	2.30E-03	3.38E-05	3.31E-02	4.94E-02
2500	AGGAGGD	MA0528.1	ZNF263	2.20E-02	1.40E-05	1.37E-02	2.74E-02
3000	CTCTSMCC	MA0504.1	NR2C2	4.10E-05	1.52E-05	1.49E-02	2.99E-02
3000	CCAYAG	MA0332.1	MET28	5.40E-03	3.94E-06	3.86E-03	7.71E-03
3000	CCAYAG	MA0335.1	MET4	5.40E-03	2.29E-05	2.24E-02	2.24E-02
3000	CGGRGA	MA0348.1	OAF1	2.20E-02	7.97E-06	7.81E-03	1.55E-02
3000	CGGRGA	MA0437.1	YPR196W	2.20E-02	4.12E-05	4.03E-02	2.87E-02
3500	CCCCTK	MA0364.1	REI1	6.50E-07	1.43E-05	1.40E-02	2.40E-02
3500	CCCCTK	MA0342.1	MSN4	6.50E-07	3.69E-05	3.61E-02	2.40E-02
3500	CCCCTK	MA0366.1	RGM1	6.50E-07	3.69E-05	3.61E-02	2.40E-02
3500	CCCCTK	MA0155.1	INSM1	6.50E-07	5.02E-05	4.91E-02	2.44E-02
3500	GCCRCA	MA0334.1	MET32	1.20E-05	4.57E-06	4.47E-03	8.95E-03
3500	GCCRCA	MA0333.1	MET31	1.20E-05	9.14E-06	8.95E-03	8.95E-03

One possible mechanism of upregulation of genes with a specific motif in their promoter region is the upregulation of an activating TF that binds to the promoter region of the gene and leads to an increase in gene expression.

The above analysis identified motifs in the promoters of upregulated genes, and identified TFs that bind to those motifs but did not examine the transcription status of these factors and their possible differential regulation in the endocardium. Given this, all the TFs identified above as matching a motif were cross-referenced with the list of differentially regulated genes.

Five transcription factors with binding sites in upregulated genes that also displayed differential expression were identified (Table 3.32). Only three of these transcription factors (Erg, Ets1 and Fli1) are upregulated. These three TFs also belong to the ETS family of transcription factors, further exemplifying the role of this family in the endocardial development. Furthermore, the match of the identified CTTCCTS motif to the ETS motifs was highly significant (Figure 3.33).

It must be noted that all the ETS family members display a highly similar motif to each other [Wei et al., 2010] and the analysis presented here may not be able to distinguish the binding of these factors on the basis of sequence similarity alone. The identified motifs may therefore bind other upregulated members of the ETS family and not one of the factors identified by the TOMTOM analysis. For this reason all the members of the ETS family that are differentially expressed in the endocardium (Table 3.33) may be considered as potential regulators of the endocardial identity.

Table 3.31: Mean number of motif occurances of each motif in the stable and upregulated set of promoters. Consistent with their identification by DREME all the motifs are more frequently identified in the upregulated set of promoter sequences.

Motif Sequences	Stable	Upregulated
AGGAGGD	7.86	10.08
AGGGGVC	3.05	3.93
ASAGGAAR	6.09	7.79
BACCCCA	3.80	5.18
BTCCCCA	4.79	6.63
CACAGMAG	3.40	4.09
CACCCTK	3.35	4.62
CACGBGC	1.06	1.14
CAGGGAS	4.69	6.28
CCCCAYCC	12.80	15.63
CCCMYTCC	7.07	9.28
CCCTGRR	2.98	4.09
CTCTSMCC	5.23	6.88
CTGRGGC	5.16	6.50
CTTCCKC	4.23	5.59
CTTCCTS	3.71	4.87
GGAAGGW	4.20	5.50
GGGTGTS	3.52	4.44
GGTGTSA	1.31	1.61
GTGACCM	2.39	3.17
HTGGGGA	4.68	6.08
RCAGGAA	2.57	3.27

Table 3.32: Transcription factors that are differentially regulated in the endocardium and their binding motif is overrepresented in the TSS of upregulated genes. All three TFs that are upregulated are also members of the ETS family of transcription factors.

Gene	$\log_2$(FC) in Endocardium
Erg	2.92
Ets1	1.07
Fli1	2.00
Tfap2a	-2.36
Tfap2c	-3.25

Table 3.33: ETS family members differentially regulated between endocardial and endothelial cells.

ETS family Gene	$\log_2$(FC) in Endocardium
Elf3	-2.87
Elk3	1.10
Erg	2.92
Ets1	1.07
Ets2	1.57
Fli1	2.00
Etv2	1.44

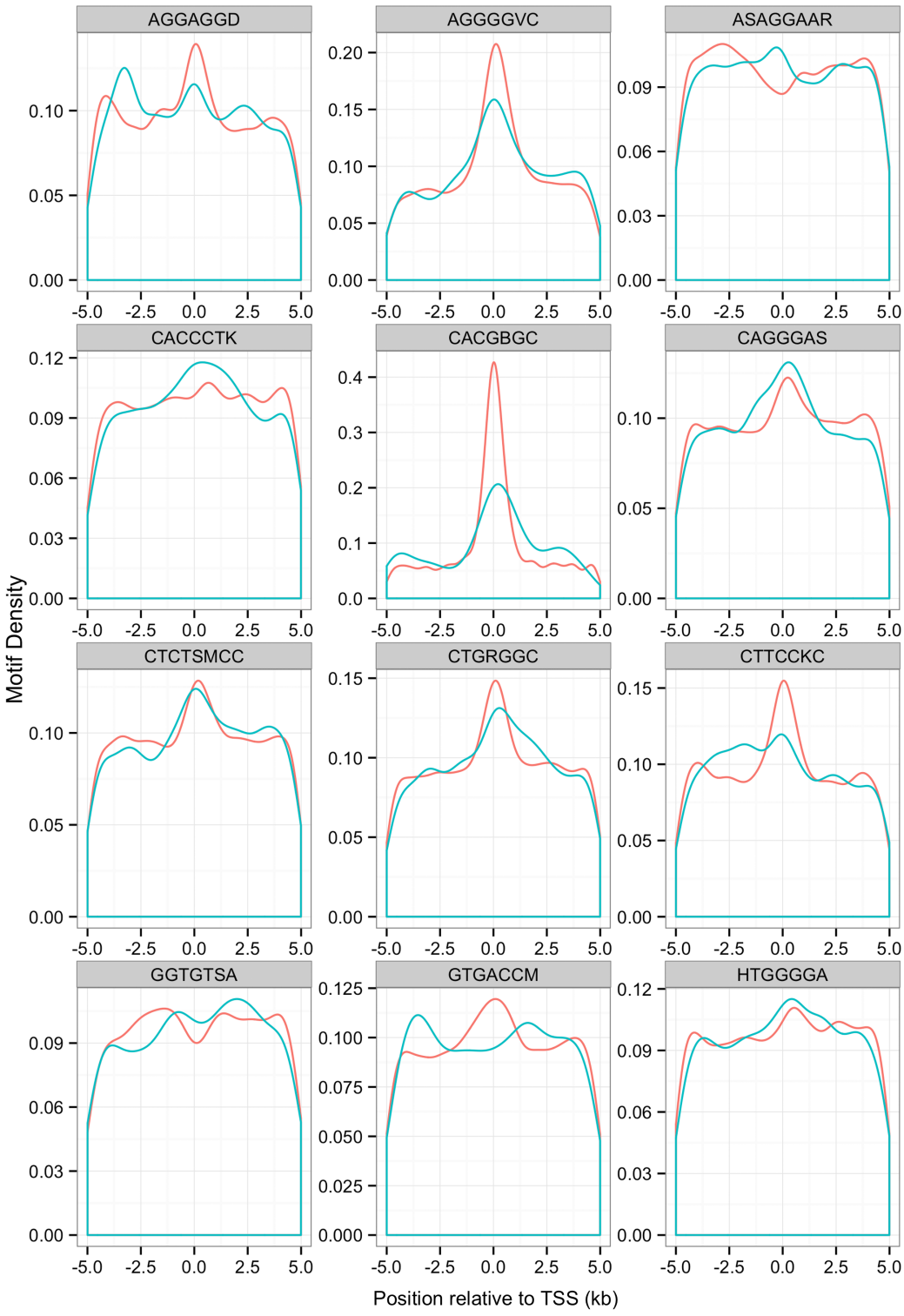


Figure 3.31: Independent identification of the distribution of motifs identified by the DREME analysis the promoters of upregulated (green) and stably (red) expressed genes reveals distinct distribution of some but not all of the motif sequences. The distributions are normalised and the height of the peak does not represent absolute abundance of the motif in each dataset. (1/2)

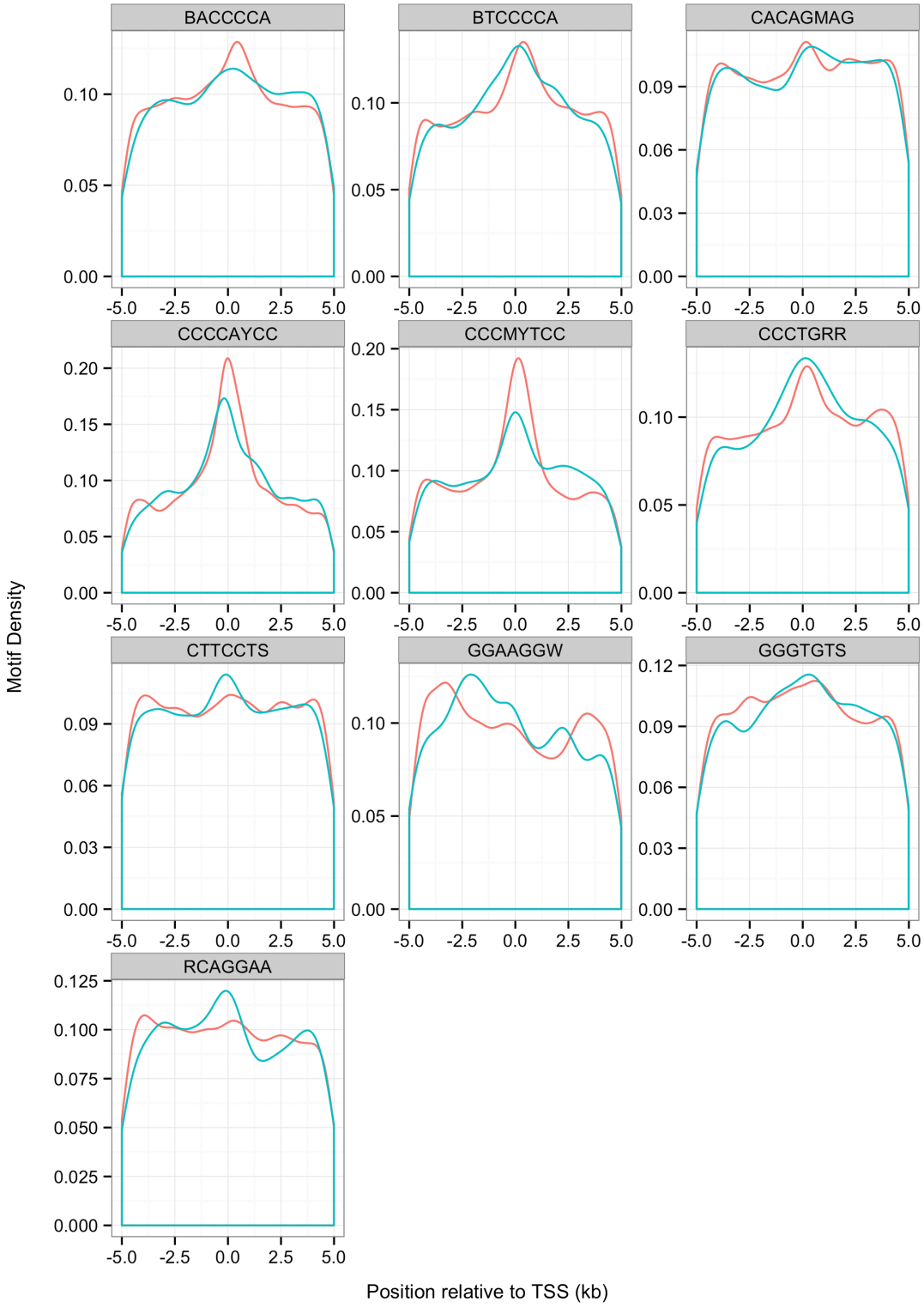


Figure 3.32: Independent identification of the distribution of motifs identified by the DREME analysis the promoters of upregulated (green) and stably (red) expressed genes reveals distinct distribution of some but not all of the motif sequences. The distributions are normalised and the height of the peak does not represent absolute abundance of the motif in each dataset. (2/2)

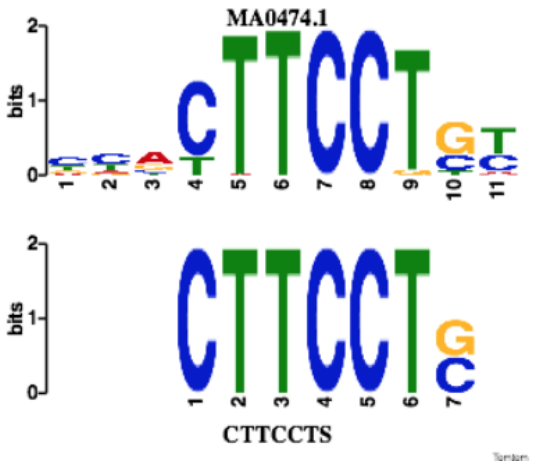


Figure 3.33: Motif sequence alignment of Erg binding motif (top) with the identified CTTCCTS motif (e-value = 3.27e-3, q-value = 3.25e-3).

Table 3.34: Significance of deviation of motif distribution from the uniform as assessed by means of the Kolmogorov-Smirnov Test for each individual motif. Multiple testing correction performed using the Bonferonni correction.

Motif	p-value	q-value
CCCCAYCC	<1.00E-10	<1.00E-10
AGGGGVC	1.27E-10	2.66E-09
GGAAGGW	3.74E-10	7.48E-09
CCCMYTCC	1.49E-07	2.83E-06
CACGBGC	7.50E-07	1.35E-05
CCCTGRR	1.95E-06	3.31E-05
CTGRGGC	1.13E-05	1.81E-04
CTCTSMCC	1.85E-05	2.77E-04
BTCCCCA	4.44E-05	6.21E-04
CTTCCKC	5.20E-05	6.76E-04
CAGGGAS	2.88E-04	3.45E-03
AGGAGGD	2.95E-04	3.24E-03
CACCCTK	1.23E-02	1.23E-01
BACCCCA	1.38E-02	1.24E-01
RCAGGAA	1.96E-02	1.57E-01
CACAGMAG	2.29E-02	1.60E-01
GGGTGTS	3.36E-02	2.01E-01
HTGGGGA	9.37E-02	4.69E-01
ASAGGAAR	9.61E-02	3.85E-01
GGTGTSA	1.48E-01	4.43E-01
GTGACCM	4.98E-01	9.96E-01
CTTCCTS	7.48E-01	7.48E-01

3.4 Comparison of Methylation and Transcriptomic Data

3.4.1 Genome-wide Relationship between Promoter CGI Methylation and Expression

The well described relationship between promoter CGI methylation status and expression was investigated in the datasets generated. For each annotated gene, the mean expression FPKM value of the gene was compared with the methylation of CGIs directly overlapping the transcription start site. The methylation status was categorised as low (<50%), or high(>50%) and the distribution of methylation for each class of CGI was plotted (Figure 3.35). The expected pattern of low methylation correlating with higher expression and vice versa was observed.

The robustness of the above result as a function of the distance between each CGI and gene pair was examined for both samples in combination and separately. The results of this analysis are shown in Figure 3.36. The top part of panel A shows the p-value of the association between the methylation status of CGIs and nearby genes as a function of distance. The association is significant for genomic distances in excess of 200 kb, however the ratio of the mean expression of genes with high and low methylation in their promoters (the overall magnitude of the effect) is reduced when the distance between the CGI and gene exceeds 10 kb and is severely diminished within 50 kb. This suggests that majority of direct interactions between methylation status and transcription state are limited by genomic distance, although interactions may persist up to 200 kb.

3.4.2 Differentially Regulated Genes Overlapping Differentially Methylated CGIs

The list of identified CGIs was cross-referenced with the list of differentially regulated genes, in order to identify genomic loci where changes in methylation coincide and may be functionally linked to control of gene expression in the context of endocardial differentiation. The robustness of this result as a function of the distance between CGIs and transcribed loci is examined in Figure 3.37. The number of overlapping loci increases approximately linearly with distance. On the basis of the analysis presented in the previous section, a distance cutoff of 50 kb was selected, as interactions beyond this point are unlikely to be direct. This analysis identifed 150 genomic locations where differential methylation and transcription colocalise and may be functionally linked. Full results of

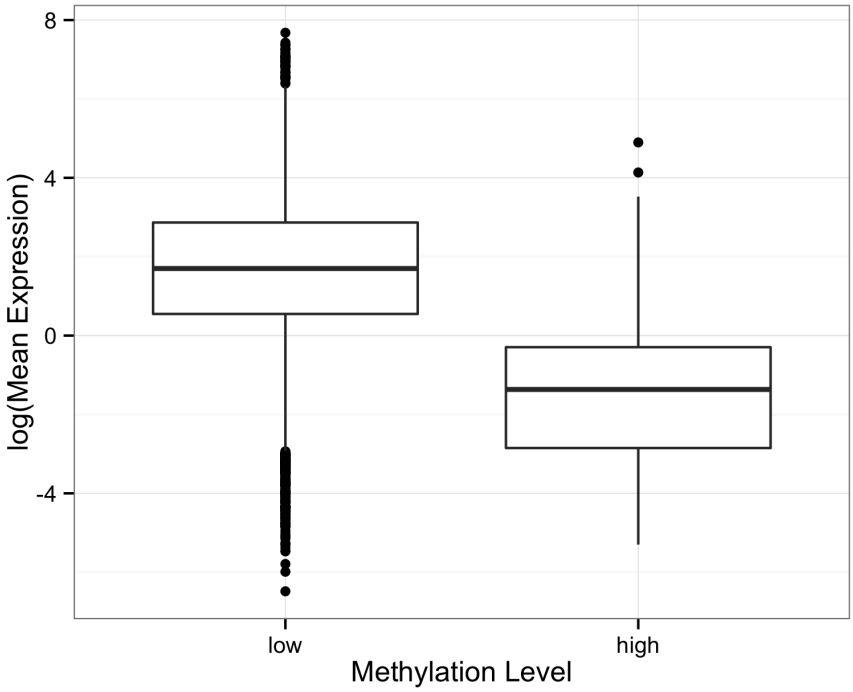


Figure 3.34: Boxplot of the mean expression of genes directly overlapping CGIs stratified by mean methylation level of CGIs, in both tissues. Genes overlapping highly methylated CGIs display significantly lower expression than genes overlapping CGIs with low methylation levels (p-value = 8.44e-29, heteroskedastic two sample T-test).

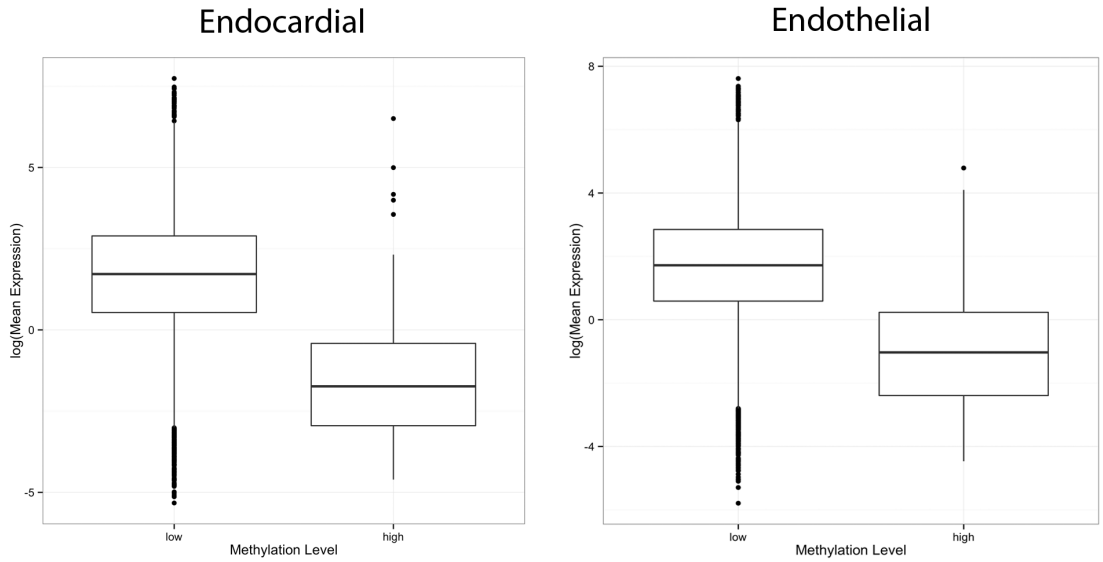


Figure 3.35: Boxplot of the mean expression of genes directly overlapping CGIs stratified by mean methylation level of CGIs, for the endocardium and the endothelium individually. Genes overlapping highly methylated CGIs display significantly lower expression than genes overlapping CGIs with low methylation levels (EC p-value = 2.305e-03, ET p-value = 4.202e-23; heteroskedastic two sample T-test).

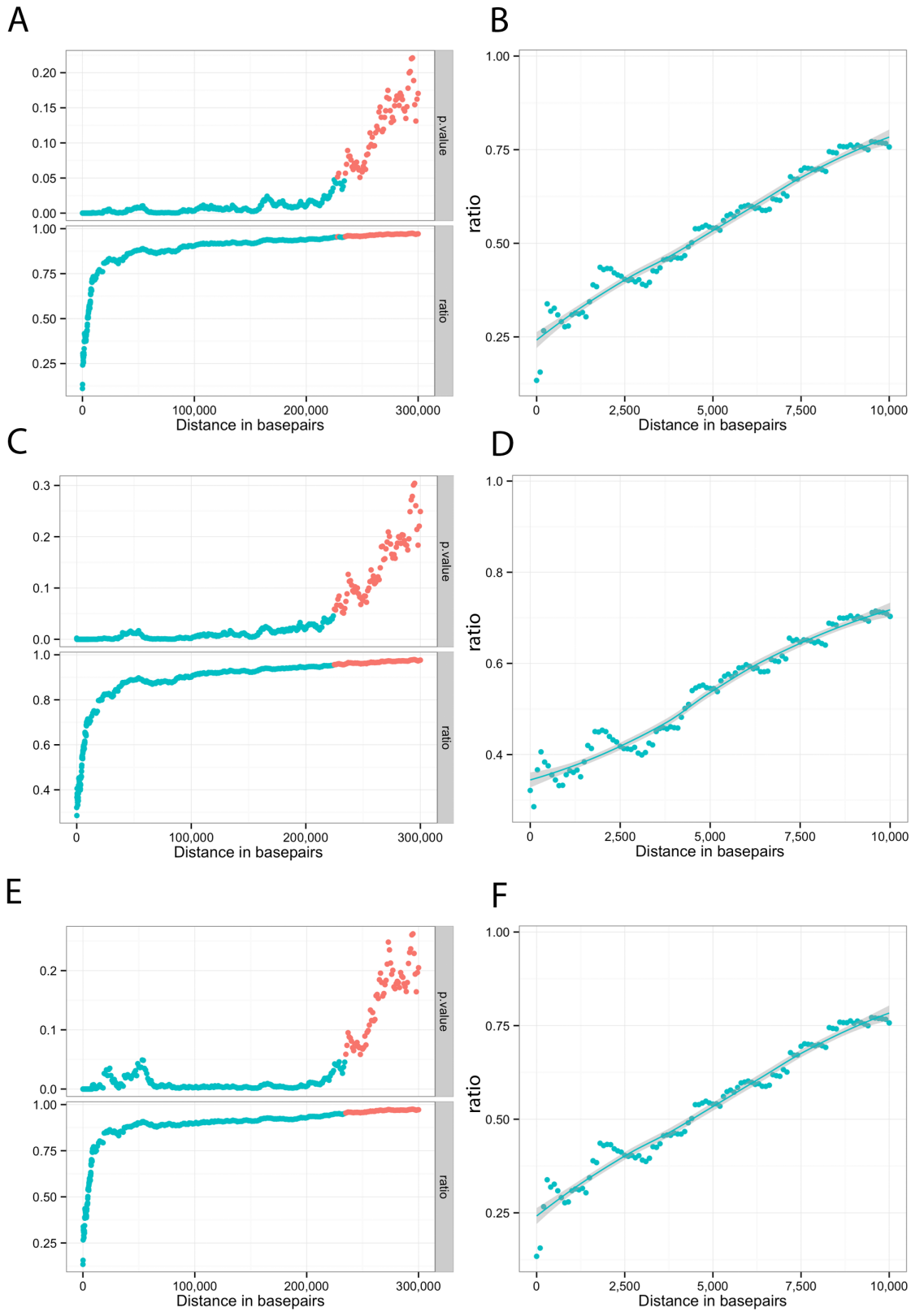


Figure 3.36: Relationship between CGI methylation and gene expression levels as a function of distance. Green denotes p-values below 0.05 and red equal to or in excess of 0.05. (A, top) p-value of association between methylation and expression as a function of distance for all data combined. The association is significant for over 200 kb. (A, bottom) ratio of the mean expression of genes overlapping low and high methylated CGIs, the magnitude of the association diminishes within 10 kb (see panel B). The relationship between CGI methylation and expression is recapitulated for Endocardial (panels C and D) and Endothelial cells (panels E and F) individually.

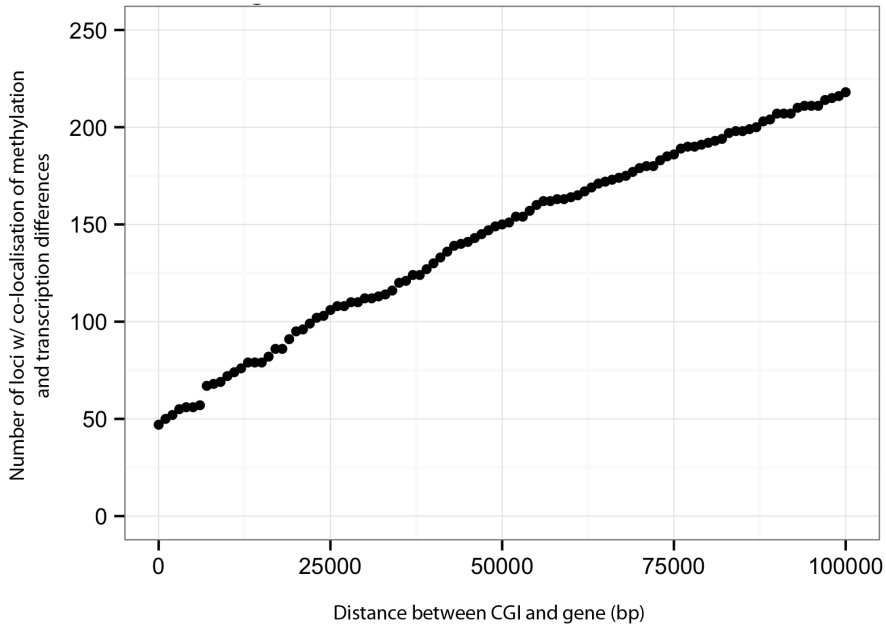


Figure 3.37: Number of times methylation and expression changes colocalise as a function of distance between respective DMRs and genes.

this analysis can be found in Table A.14 of the Appendix.

It is of interest to note that several genes of known importance appear in the results of this analysis. These include *Tal1*, which among others is critical for regulation of intercellular junction in the endocardium, *Tie1* a tyrosine kinase receptor which is critical for the development of the endocardium as well as *Elf3* a member of the ETS family of transcription factors that are identified in Section 3.3.9 as potential regulators of the endocardial identity.

3.4.3 Differentially Regulated Genes Overlapping DMRs

Further to the above, the list of differentially methylated regions identified genome-wide was cross-referenced to the list of differentially regulated genes with the aim of identifying genomic loci where methylation and gene expression are potentially functionally linked. Similarly to the above analysis the cut-off distance for the two changes to be considered to be colocalised was set to 50 kb.

The analysis revealed that 107 DMRs colocalise with 97 differentially expressed genes between the endocardium and the endothelium. Hypermethylation is weakly anticorrelated with expression (Spearman ρ = -0.22). The results of this analysis can be found in Table 3.35. Similar to the analysis reported above, DMRs in the vicinity of known endocardial regulators such as *Tal1* and *Tie1* were identified, further supporting the biological

Table 3.35: Overlaps of identified DMRs with differentially regulated genes, within a distance cutoff of 50 kb. This analysis identified some well established regulators of endocardial identity such as *Tal1* and *Tie1* as potentially epigenetically regulated. Some genes appear multiple times as they can have multiple isoforms or overlap more than one DMR.

Gene	DMR Locus	EC methylation	ET methylation	EC expression	ET expression
Novel Transcript	chr7:20055181-20055460	73.0%	43.9%	5.02	1.75
Novel Transcript	chr8:124793057-124793175	72.1%	44.4%	1.80	3.24
Novel Transcript	chr8:124820289-124820793	30.6%	51.1%	1.80	3.24
Abca4	chr3:121787724-121788914	68.6%	40.8%	1.80	3.28
Ablim1	chr19:57064818-57065598	78.4%	47.2%	15.33	10.01
Adam11	chr11:102639062-102639563	65.8%	42.8%	0.78	1.98
Adcy4	chr14:56380507-56380930	40.6%	16.1%	3.60	1.19
Adcy4	chr14:56434090-56434667	73.2%	49.2%	3.60	1.19
Adora2a	chr10:74816185-74817404	58.1%	80.1%	4.71	0.93
Alox5ap	chr5:150057323-150058420	35.9%	59.0%	13.46	3.78
Ano9	chr7:148274750-148274879	78.4%	52.4%	0.31	2.86
Aqp3	chr4:41044457-41044778	73.1%	46.3%	0.86	4.42
B4galnt4	chr7:148211594-148212613	72.1%	39.5%	8.56	15.26
B4galnt4	chr7:148274750-148274879	78.4%	52.4%	8.56	15.26
BC068157	chr8:4206269-4206850	55.2%	23.4%	0.61	2.88
Bcl6b	chr11:70028833-70029223	53.3%	27.6%	4.24	0.90
Capn5	chr7:105313648-105314322	77.3%	49.8%	15.25	7.23
Ccm2l	chr2:152896544-152896824	43.1%	23.0%	2.85	0.34
Celsr2	chr3:108187823-108188389	83.4%	52.8%	1.98	5.90
Cep170b	chr12:113958407-113959505	65.0%	34.8%	6.79	10.60
Cldn3	chr5:135442964-135443614	78.9%	50.9%	0.59	3.61
Cldn4	chr5:135442964-135443614	78.9%	50.9%	3.95	19.31
Clic6	chr16:92475463-92476098	55.5%	32.4%	9.27	3.22

Clic6	chr16:92468411-92468768	80.5%	58.2%	9.27	3.22
Col8a2	chr4:126002087-126002467	74.6%	45.8%	8.90	15.02
Col8a2	chr4:125923594-125924274	82.2%	59.7%	8.90	15.02
Csf1r	chr18:61228464-61228781	76.0%	52.7%	10.79	4.20
Csf1r	chr18:61225022-61225835	49.3%	26.6%	10.79	4.20
Cyp26a1	chr19:37765321-37765560	34.0%	11.5%	19.22	10.68
Cyp26a1	chr19:37762177-37762475	52.9%	32.7%	19.22	10.68
Cyp26a1	chr19:37814091-37814949	56.4%	77.6%	19.22	10.68
Dbx1	chr7:56887804-56888417	65.3%	32.0%	0.51	3.88
Ddah2	chr17:35154731-35154891	80.8%	58.8%	158.03	108.21
Dll1	chr17:15519420-15519960	80.6%	59.4%	1.08	3.65
Dnaaf3	chr7:4434397-4434994	86.5%	62.6%	3.67	8.86
Dsg2	chr18:20759818-20760284	84.9%	53.9%	7.66	11.89
Dst	chr1:34093309-34093859	84.1%	61.2%	9.80	16.06
Ehbp1l1	chr19:5751116-5751276	40.8%	19.9%	10.02	6.10
Elf3	chr1:137114015-137114765	71.5%	42.4%	0.66	4.83
Epb4.1	chr4:131478296-131478516	62.6%	42.4%	37.69	25.74
Esrrb	chr12:87720288-87721157	77.9%	43.7%	0.65	6.28
Exoc3l	chr8:107860607-107860809	73.6%	44.9%	4.72	1.48
Ezr	chr17:6940931-6941564	74.8%	50.2%	55.86	79.12
Ezr	chr17:6988252-6988957	80.4%	59.1%	55.86	79.12
Fam198b	chr3:79683120-79683489	84.2%	58.3%	2.95	0.97
Fam65a	chr8:108090552-108090881	50.2%	20.4%	37.58	20.41
Flt4	chr11:49403109-49403751	85.9%	64.2%	14.91	3.56
Foxa2	chr2:147849058-147849438	73.3%	48.4%	0.96	3.54
Fzd10	chr5:129099014-129099686	78.1%	53.3%	13.82	8.78
Gata2	chr6:88154810-88156434	62.6%	41.6%	8.45	2.07
Gdf3	chr6:122513110-122513678	79.8%	57.5%	0.28	4.04
Grap2	chr15:80416857-80417546	78.6%	55.3%	10.22	5.09
Grik3	chr4:125212061-125213486	50.7%	15.6%	0.68	2.47
Hhex	chr19:37507291-37507847	76.1%	53.9%	15.41	4.11

Ifi30	chr8:73269722-73269943	86.8%	54.0%	38.63	60.68
Ifi30	chr8:73309871-73310263	80.4%	53.4%	38.63	60.68
Ifi30	chr8:73334834-73335143	76.8%	49.9%	38.63	60.68
Ifitm1	chr7:148192086-148193168	88.7%	62.9%	49.99	92.19
Igf1	chr10:87323389-87323840	57.1%	32.1%	8.47	1.78
Igfbpl1	chr4:45801580-45802815	58.7%	79.1%	0.68	2.82
Lefty1	chr1:182824490-182825286	54.9%	27.8%	1.09	4.65
Lfng	chr5:141088065-141088599	55.9%	34.7%	2.40	5.90
Lmo2	chr2:103843358-103843795	70.0%	38.1%	26.72	6.16
Lmo2	chr2:103843358-103843795	70.0%	38.1%	3.42	1.04
Ltbp4	chr7:28066903-28067263	76.2%	50.4%	5.45	11.82
Lyve1	chr7:117952877-117953693	86.9%	66.2%	1.68	0.23
Map7	chr10:19861332-19862037	80.3%	50.3%	1.59	4.47
Msi1	chr5:115894491-115895052	88.4%	61.2%	23.36	42.97
Myh14	chr7:51857056-51858580	52.7%	25.7%	0.49	1.32
Nr5a2	chr1:138696965-138697319	91.0%	69.4%	0.40	1.46
Pecam1	chr11:106468008-106468818	81.0%	50.9%	52.41	28.46
Plxnd1	chr6:115912422-115912546	84.0%	62.7%	42.55	13.29
Polg	chr7:86503189-86504238	35.9%	63.3%	83.43	52.70
Pou3f1	chr4:124344632-124345709	61.9%	37.0%	0.40	1.43
Ppp1r13b	chr12:113027316-113027845	84.6%	55.8%	14.51	8.75
Ptrf	chr11:100819860-100820506	45.7%	25.5%	26.92	18.03
Rasip1	chr7:52887987-52888323	68.3%	42.7%	12.56	1.84
Rcsd1	chr1:167565599-167566481	72.5%	48.5%	20.12	10.46
Sesn3	chr9:14050690-14051104	61.9%	38.8%	6.57	10.60
Sh2b3	chr5:122278315-122278607	62.6%	35.1%	17.13	8.22
Sh3tc1	chr5:36083518-36083985	47.7%	21.4%	3.92	1.83
Skap1	chr11:96338361-96338945	90.2%	66.9%	5.66	0.84
Slc30a10	chr1:187308178-187308821	88.1%	65.2%	4.53	1.83
Slc4a1	chr11:102217890-102218476	79.4%	57.5%	1.60	4.06
Slc7a8	chr14:55330531-55330769	87.5%	62.5%	4.96	2.23

Smad6	chr9:63801417-63801870	80.0%	41.3%	22.69	15.14
Smad6	chr9:63775068-63775725	50.0%	24.7%	22.69	15.14
Smad6	chr9:63766729-63767498	53.8%	74.2%	22.69	15.14
Snhg11	chr2:158222992-158223588	87.9%	67.0%	13.70	21.40
Spint2	chr7:29997949-29998493	63.5%	39.3%	7.60	20.60
Spock2	chr10:59570816-59571512	58.2%	36.8%	1.23	4.31
Tal1	chr4:114751244-114752324	65.2%	42.3%	31.18	6.31
Tal1	chr4:114751244-114752324	65.2%	42.3%	6.20	1.32
Tal1	chr4:114751244-114752324	65.2%	42.3%	5.76	1.21
Tcea3	chr4:135846684-135847192	49.2%	17.6%	0.73	4.96
Tdgf1	chr9:110817103-110817617	80.1%	48.3%	3.34	14.76
Tgfb1	chr7:26430147-26430915	75.7%	50.7%	41.90	27.61
Tgfb1	chr7:26489381-26490333	38.0%	13.9%	41.90	27.61
Tie1	chr4:118108001-118108883	81.0%	52.6%	40.45	6.62
Tie1	chr4:118190073-118190950	61.8%	33.6%	40.45	6.62
Tinagl1	chr4:129807866-129808438	73.3%	51.9%	27.21	41.91
Tinagl1	chr4:129825701-129826354	87.5%	66.4%	27.21	41.91
Tnfaip2	chr12:112661520-112662384	40.5%	20.5%	24.62	7.49
Tnfrsf19	chr14:61623709-61624298	38.0%	13.2%	9.47	17.47
Ushbp1	chr8:73928811-73929522	63.0%	37.7%	2.22	0.35
Wnt1	chr15:98626660-98627388	60.9%	39.3%	0.64	5.51
Wnt7b	chr15:85337985-85338351	81.0%	58.9%	0.26	2.20
Wwc1	chr11:35610726-35611148	73.5%	51.7%	3.52	8.43
Zfpm1	chr8:124793057-124793175	72.1%	44.4%	28.72	14.06
Zfpm1	chr8:124820289-124820793	30.6%	51.1%	28.72	14.06
Zfpm1	chr8:124849801-124850807	24.3%	45.6%	28.72	14.06
Zfpm1	chr8:124845529-124845986	35.3%	57.8%	28.72	14.06
Zic2	chr14:122852090-122852827	69.2%	43.8%	1.97	5.17
Zswim5	chr4:116502237-116503120	62.7%	38.4%	0.80	2.00

relevance of methylation in these loci.

3.5 Discussion

The endocardium is an epithelial layer of cells lining the lumen of the ventricles and atria in the developing and adult heart. This cell population has significant roles in trabeculation of the myocardium, differentiation of cardiomyocytes into conduction fibers, EMT at sites of valve formation and septation of the OFT into the pulmonary artery and aorta [Harris and Black, 2010]. More recently the endocardium has been found to be the source of part of the coronary vasculature [Tian et al., 2014].

The endocardium first appears after gastrulation at the same time as the formation of the cardiac crescent, the primitive structure that will give rise to the adult heart [Baldwin, 1996] [Gilbert, 2003] [Abu-Issa and Kirby, 2007]. The origin of the endocardium remains unclear however, with conflicting reports in the literature of a vascular or cardiac origin.

NFATc1 has been identified as a transient marker of the endocardium that is exclusively expressed in all of the endocardium between E8.5 and E9.5 [Ranger et al., 1998] [de la Pompa et al., 1998]. Shortly after E9.5 NFATc1 expression pattern is altered and it ceases to be expressed in the entirety of the endocardium. At this point NFATc1 is downregulated in the ventricular endocardium and upregulated at the sites of valve formation, where it has a role in the control of EMT [Wu et al., 2013] [Zhu et al., 2013]. A transgenic construct, utilising the unique expression pattern of NFATc1, allows identification of endocardial cells *in vivo* and in culture [Misfeldt et al., 2009].

Endocardial development has been previously shown to be recapitulated during embryoid body differentiation [Narumiya et al., 2007] [Misfeldt et al., 2009]. Isolation of endocardial cells from embryoid bodies has the significant advantage that hundreds of thousands of cells can be effectively isolated. Isolation of similar numbers of cells from embryos is technically challenging.

In the context of this work, embryoid body derived endocardial cells were compared with embryoid body derived endothelial cells in order to identify unique epigenetic and transcriptional regulators of this cardiac cell line. Cells were isolated from embryoid bodies using the NFATc1 marker outlined above and the CD31 surface molecule unique to the endothelium. Endocardial cells were identified by the double positive signature NFATc1^{+}/CD31^{+}, whereas endothelial cells were identified as NFATc1^{-}/CD31^{+}.

It was hypothesised that epigenetic processes, and genomic methylation in particular, contribute strongly to the identity of the endocardium and are a determining factor in its differentiation from other endothelium. This hypothesis was based on the identical morphology of the endocardium and endothelium, as well as prior microarray and RNA-seq differential expression analysis that suggested that no differentially regulated genes existed between the two cell types. A plethora of prior evidence for the role of epigenetic processes in heart development further supported this hypothesis (Section 1.3.6).

Whole genome-methylation analysis was performed on endocardial and endothelial cells. A library preparation protocol for WGBS was developed using the RRBS protocol as a template. The protocol was used to prepare WGBS libraries from endocardial and endothelial cells. A pipeline for the processing of the resulting data and for the identification of differentially methylated CGIs and genomic DMRs was developed.

A similar but not identical methylome profile was identified in the two cell types examined. This is consistent with a common developmental origin for endocardial and endothelial cells, suggesting a vascular precursor for the endocardium. However, the present study did not examine the methylation profiles of cardiac progenitors, so the proximity of the identified profile to those progenitors could not be determined. It is therefore impossible to evaluate if the endocardial methylation profile is more similar to the endothelial methylation profile or that of the early heart precursors.

Differential methylation analysis was performed at the CGI level and at the whole genome level using a smoothing algorithm. CGI methylation analysis revealed 1,641 differentially methylated islands between the two cell types at the $p < 0.05$ level of significance. However, none of these islands were significantly differentially methylated after multiple testing correction and their distribution did not appreciably depart from the distribution of all examined CGIs. Subsequent, whole genome analysis identified a 1,128 windows of differential methylation that were preferentially enriched at functional genomic sites and enriched in the vicinity of transcribed genes related to developmental processes.

The analysis of the above cell populations was extended to an analysis of their transcriptome. As aforementioned, past microarray and transcriptomics analyses by others had proved to be unfruitful. With an expectation for low effect sizes mRNA-seq was performed on independent quadruplicate pools of endocardial and endothelial cells. In sharp contrast to past results, this analysis identified a large number of differentially regulated

genes, including genes already implicated in the endocardial identity.

Gene Ontology (GO) overrepresentation analysis was performed on the list of differentially regulated genes in aggregate and stratified by cell type of upregulation. Results of the aggregate analysis were consistent with the endocardial and endothelial identity of the population compared. Interestingly, the GO term analysis suggested differential regulation of genes that respond to the local hemodynamic environment. The stratified analysis suggested an enrichment of both vascular and haematopoietic specifically in the endocardium and confirmed that genes related to response to fluid shear stress are specifically overrepresented in the endocardium.

In an effort to identify factors that may have an upstream role in specification and maintenance of the endocardial identity, the transcription start sites of the upregulated genes were search for the overrepresentation of sequence motifs. After stringent testing criteria 22 motifs were identified that closely resembled 45 known motif sequences.

In order to prioritise by likely biological significance, the identified motifs were mapped to cellular factors that are known to bind them and the list of the obtained cellular factors was cross-referenced to the list of differentially regulated genes. This analysis was expected to identify differentially regulated factors that directly bind to the promoters of and influence the expression of other genes. Five differentially regulated transcription factors were found to match overrepresented motifs, of which the only upregulated ones were members of the ETS family of transcription factors. Members of the ETS family of transcription factors are known to play a role in vascular differentiation and the present analysis suggests that they may have a more specific role in endocardial differentiation.

Comparison of the methylome and transcriptome data generated, revealed the expected pattern of promoter hyper-methylation being associated with lower expression of genes, suggesting that the methylome profiles of the cells are consistent with the transcriptome profiles generated. This pattern was observed independently for endocardial and endothelial cells and persisted when up to 200 kb separated the two elements. Specific comparison of the list of differentially regulated transcripts and differentially regulated CGIs identified 150 genomic locations that show alteration in both methylation and expression status. Further comparison of the list of the identified DMRs with differentially expressed genes identified 97 genes that overlap DMRs, some of which are known to have a role in endocardial development. This establishes a potential direct link between endocardial development and differential methylation in this tissue.

3.5.1 Limitations of the Present Study

A significant limitation of the design of this study is the analysis of endocardial cells past the time point of specification. Although the exact time point of specification of the endocardial cell line is not known, it can be reasonably hypothesised to occur prior to the time of the first appearance of the endocardium by downregulation of N-cadherin. This event precedes E8.5, the earliest time point of NFATc1 expression. This limitation is significant because there is no certainty that traces of the early molecular events that result in specification will persist at E9.5. This limitation is however not currently surmountable as no markers for the endocardium prior to NFATc1 are known. Only single cell analysis (see following Section) combined with micro-dissection could potentially overcome this limitation without the discovery of an earlier marker. Such an investigation however is extremely technically demanding.

In addition, another limitation of the work presented here is that the whole-genome bisulphite sequencing assay employed can not distinguish between hydroxymethylation and methylation of DNA. Consequently the level of both DNA modifications is measured. This limitation of the bisulphite conversion reaction is not relevant in most tissues as the levels of hydroxylmethylation are low. Given however that the endocardial cells examined are obtained by differentiation of the embryonic stem cells, in which hydroxylmethylation is known to play a significant role, the inability to distinguish between the two cytosine modification states may hamper the analysis. It is however expected that such influence will be small given that differentiation of embryonic stem cells for approximately seven days has occurred *in vitro* and the cells examined are morphologically distinct from the embryonic cells used for differentiation. This limitation can be overcome, at a signifcant financial cost, by oxidative bisulphite sequencing which can distinguish between the two DNA modifications [Booth et al., 2013].

Despite the above, the analysis presented here is important because it establishes that transcriptomic and epigenetic differences exist between the two cell lines, justifying a more challenging investigation. Furthermore, the present study provides us with an array of candidates that upon further investigation may include currently unknown earlier endocardial markers and guide further analyses.

In summary, the work presented here has for the first time provided us with comprehensive epigenomic and transcriptomic profiles for endocardial and endothelial cells. The differential expression analysis has identified a plethora of differentially regulated genes

between the two cell types, providing us with an expansive list of candidates for further characterisation. Differential methylation analysis has provided us with a list of 1,128 candidate regions of between the two cell types. The functional significance of these sites is supported by their overlap with epigenetic marks with known regulatory functions, as well as being in the vicinity of genes associated with developmental processes. Bioinformatic analysis was used to identify specific members of the ETS family of transcription factors as likely regulators of a large number of the genes differentially expressed.

3.5.2 Further Work

The results of the genome-wide analyses presented here should be validated using locus specific qPCR assays prior to further investigation. Differentially regulated genes should be prioritised for validation using functional criteria, such as transcription factor activity and lists of these genes are presented in the context of this work. Differentially methylated regions should also be validated using locus specific assays. Prioritisation of these regions is more challenging but DMRs that colocalise with genes known to be relevant to endocardial development, such as *Tal1* constitute important targets in addition to the regions that display maximal methylation changes.

The model of differentiation used here has previously been shown to recapitulate embryonic development to a significant extent, however differences to the *in vivo* differentiation process can be expected. It would thus be of value to examine the spatial and temporal expression of the differentially expressed genes identified here via immunohistochemistry and *in situ* hybridisation in E9.5 and earlier mouse embryos. This analysis is expected to provide more clear evidence on the biological relevance of these hits. Furthermore, examination of earlier time points has the potential to identify some of the differentially expressed genes as novel early markers of the endocardium.

The applicability of the embryoid body differentiation model in the context of endocardiogenesis has been previously assessed by examination of selected marker genes and the spatial and temporal pattern of endocardial growth [Misfeldt et al., 2009]. It would be of particular interest to compare the transcriptomic profiles of *in vivo* endocardial cells and endothelial cells. This comparison has the potential to reaffirm the relevance of the model and can point to differences between the two cell types. Efforts to complete this type of analysis in the context of the project were hampered by technical challenges in isolating the required number of cells from E9.5 embryo hearts and making the mRNA-seq

libraries, within the time constraints of this project.

The above difficulties could be overcome by the application of single cell mRNA-seq techniques. Furthermore, the application of single cell technologies has the potential to overcome the challenges associated with heterogeneous populations of cells. As aforementioned, the endocardium is a heterogeneous population as the ventricular portion has clearly distinct properties from that of valvular endocardium at later developmental stages. The endothelial population examined is also highly diverse. Application of single cell mRNA-seq to cells isolated from embryos would allow unprecedented spatial and temporal resolution, allowing the identification of regulators of specific events (such as EMT and trabeculation) in the development of the endocardium and assessment of the homogeneity of the endocardium at different time points. A recent study has demonstrated the feasibility of this approach in the elucidation of alveolar progenitor cells [Treutlein et al., 2014].

Analysis of other epigenetic events beyond methylation continues to be of particular interest. In particular, assessment of the - less stable and more closely related to transcriptional regulation - histone modifications via ChIP-seq would be of interest. These analyses are however very technically challenging with low number of cells and will not be applicable to cells from mouse embryos. Some progress in the isolation of chromatin from low numbers of cells was reported by the author, but it was not feasible to obtain sufficient DNA after immuno-precipitation within the time limits of this project.

In addition to the examination of histone modifications via ChIP-seq, the identification of binding sites of the transcription factors identified as upregulated in the context of this project would be of interest, this is especially true for transcription factors in the ETS family. Finally, efforts to identify the binding sites of NFATc1 in endocardial cells are under way.

3.6 Conclusion

In conclusion, in an effort to identify factors that are responsible for the endocardial identity, the work presented here compared the methylome and transcriptome of endocardial and endothelial cells. The study identified 1,128 differentially methylated genomic locations between the two cell types. Transcriptome analysis revealed extensive expression changes between the two cell types and identified 711 differentially regulated genes.

GO term analysis confirmed that the expected biological processes were overrepresented in the list of differentially regulated genes. Interestingly, the overrepresented GO terms suggest that the local haemodynamic environment may have a contribution to the endocardial phenotype. Motif analysis of the promoter regions of differentially regulated genes, suggested a previously unappreciated role for ETS family members in endocardial development, seven members of which show differential regulation between the endocardium and the endothelium.

Examination of the methylome and transcriptome data in concert, confirmed the known relationship between promoter CGI and gene body expression in our dataset and identified 150 loci where the methylation pattern is altered in concert with gene expression.

The present study has provided valuable evidence suggesting that both methylome and transcriptome differences between endocardial and endothelial cells have a role in the specification of the endocardium and has provided us with a plethora of candidates for the regulation of the endocardial identity in anticipation of further *in vivo* experimental investigation.

Chapter 4

Allele-specific CTCF and cohesin Binding in the Mouse Brain

Part of the work presented in this chapter has been published previously in [Prickett et al., 2013] and selected Figures and Tables are reproduced in the context of this work. Unless noted, the author has had a major contribution to or prepared in their entirety all the Figures and Tables reproduced here.

CTCF is an 11 Zn-finger DNA binding factor first characterised by Filippova and colleagues [Filippova et al., 1996]. CTCF is usually described as an insulator protein, although roles in transcriptional regulation and nuclear organisation have emerged (Section 1.5). CTCF is known to bind unmethylated DNA preferentially and this preference is responsible for its well characterised regulation of the *H19* imprinted locus, where it is implicated to the allele-specific expression of the *H19* and *Igf2* transcripts (1.5.3). CTCF is therefore known to be part of a mechanism that mediates the conversion of underlying epigenetic signals (in this case methylation) to expression differences. CTCF has also been directly proposed to be part of a wider system of heritable epigenetic regulation [Phillips and Corces, 2009], although this has not been demonstrated experimentally.

Cohesin is a protein complex with a well described role in sister chromatin cohesion and their segregation during anaphase (Section 1.6). In addition, cohesin is known to co-localise with CTCF during interphase and has furthermore been implicated in transcriptional regulation in the Interferon-γ locus [Hadjur et al., 2009].

Given the close association of CTCF and cohesin, the well characterised role of CTCF in the regulation of imprinting at the *H19* locus, the known role of regulation of expression by cohesin and the proposed action of CTCF as part of an heritable epigenetic mechanism,

we performed allele-specific ChIP-seq on post-natal day 21 (P21) mouse brain. Brain is a tissue in which imprinting is known to occur extensively and of the 104 imprinted transcripts known in the mouse [Schulz et al., 2008], more than 50 are known to be expressed in the brain [Wilkins, 2008].

4.1 ChIP-seq for Detection of Parent-of-origin Specific Binding of CTCF and cohesin

4.1.1 Interspecies Crosses, ChIP-seq and Sequencing

Reciprocal crosses between C57BL6J (Bl6) and *castaneous* mouse strains were performed and F_1 progeny P21 brain tissue was banked before this project. The experimental design is shown diagrammatically in Figure 4.1. BxC animals derive from a cross where the dam is Bl6 and the sire *castaneous* and CxB animals from crosses where the sire is Bl6 and the dam *castaneous*.

Reciprocal crosses were used to differentiate between allele-specific DNA binding in a sequence dependent manner from true parent-of-origin specific binding and furthermore ensured that mapping bias to the reference sequence (Bl6) was not influencing the results (Section 4.6.4).

ChIP-seq for CTCF and cohesin was performed separately by Dr Adam Prickett on the above tissues as described in Section 2.8.1. In total, four sequencing libraries were prepared and sequenced, two corresponding to the CTCF ChIP-seq and two to the cohesin ChIP-seq.

Paired-end sequencing of the libraries was performed on four lanes of an Illumina GAIIx instrument by the BRC Genomics Facility. One instrument lane per library was used for sequencing. Details of the libraries sequenced and raw read counts for the ChIP experiments and the input can be found in Table 4.1.

4.1.2 Primary ChIP-seq Data Analysis

Quality control of the reads and alignment to the mm9 reference genome was performed by Dr Reiner Schulz as detailed in Section 2.8.2. Duplicate reads were identified and removed using the Picard toolkit. Percent duplication rates are shown in Figure 4.2. The level of duplication was low and conducive to further analysis.

Table 4.1: CTCF and cohesin ChIP-seq library read counts.

Library	Read Count	Duplicate Reads
CTCF_BxC	118,102,950	3,659,373
CTCF_CxB	126,716,090	6,589,034
Rad21_BxC	130,812,978	7,572,405
Rad21_CxB	115,163,984	7,207,008
input_CTCF_BxC	33,786,908	8,881,349
input_CTCF_CxB	32,960,735	1,907,527
input_Rad21_BxC	33,843,811	2,961,489
input_Rad21_CxB	34,237,535	2,287,633

Figure 4.1: Summary of experimental design. ChIP-seq was performed for CTCF and the rad21 cohesin subunit on P21 brain in BxC and CxB F_1 hybrid animals. Adapted from [Prickett et al., 2013]. Figure prepared by Dr Adam Prickett.

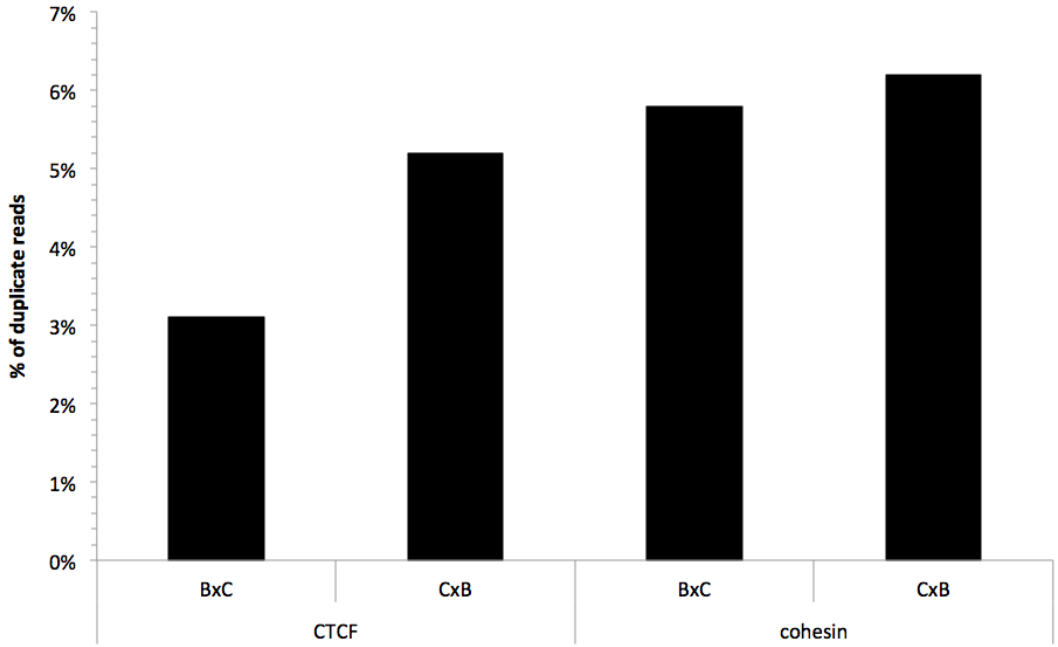


Figure 4.2: Duplication rate across all examined libraries. Duplication rate low and was less than 10% for all libraries. Adapted from [Prickett et al., 2013].

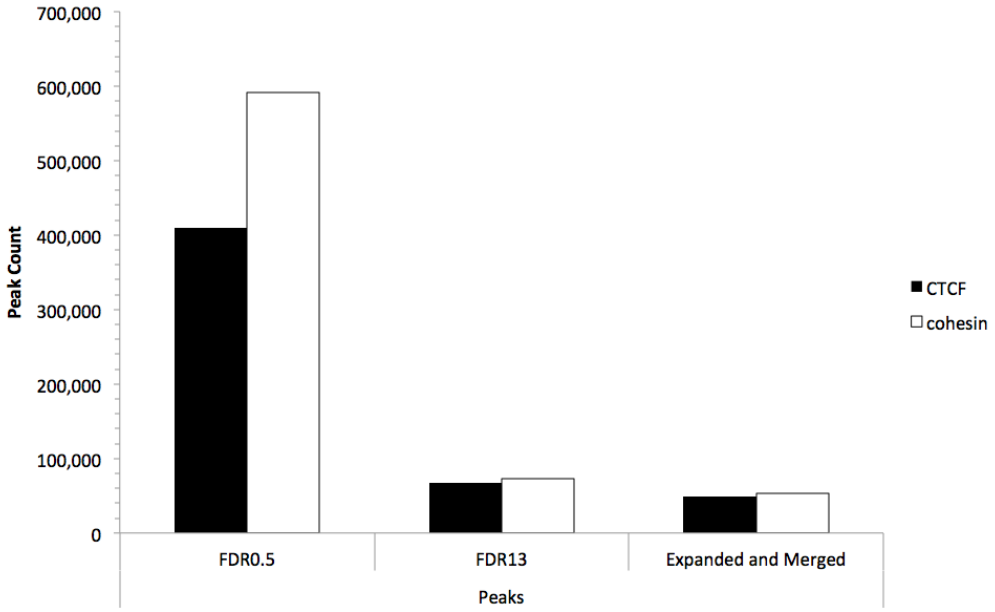


Figure 4.3: Bar plot of the number of CTCF and cohesin peaks after initial identification at FDR 0.5, refinement to FDR 13 and expansion and merging. Expansion and merging of peaks has a minor effect on the total count of peaks.

4.1.3 Identification of CTCF and cohesin Binding Sites

Identification of CTCF and cohesin binding sites with USeq was performed prior to this work. Peaks were initially detected as detailed in Section 2.8.2 to a very permissive cutoff of Phred scaled False Discovery Ratio 0.5 (3% accuracy). This was performed by Dr Reiner Schulz, all subsequent work was performed by the author.

The set of peaks was refined to include only peaks of FDR 13 or higher (95% accuracy) by filtering using a custom UNIX script to discard peaks unlikely to be biologically relevant. Peaks in the resulting set were expanded by +/-500bp and overlapping peaks were merged.

Expansion and merging of overlapping peaks was not essential at this processing stage, however it was important for subsequent analysis (Section 4.6.2) and was performed here for consistency. The number of peaks throughout the processing pipeline is shown in Figure 4.3. Out of the approximately 70,000 peaks in each dataset, 20,000 peaks were removed by merging.

4.1.4 Overlap of CTCF and cohesin Binding Sites

CTCF and cohesin have been previously reported to co-localise, although independent roles have also been reported. In order to confirm this observation, we assessed the overlap of the detected peaks.

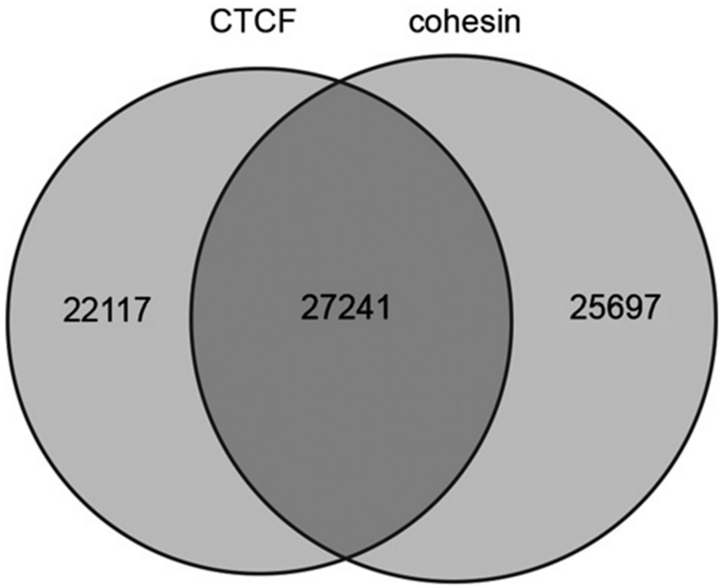


Figure 4.4: Venn diagram of overlap of CTCF and cohesin binding sites in the mouse brain. Approximately half of their binding sites are shared (55% of CTCF and 51% of cohesin), supporting concerted and independent action for both proteins. Adapted from [Prickett et al., 2013].

Although an intuitively simple concept, peak overlap between two datasets is not trivial to calculate because the number of overlaps between two sets of genomic intervals A and B can either refer to the number of elements in A overlapping one of more elements in B or vice versa. These two numbers are not by definition equal and may in fact differ considerably. For the context of all subsequent work, peak overlaps were calculated as the mean of the two reciprocal overlap calculations. For this reason, the sum of the number of peaks may not always accurately recapitulate the total sum of peaks reported. Peaks were reported as overlapping when an overlap of at least 1 bp between the two peaks existed.

CTCF and cohesin peaks show considerable overlap and share between 51.4% and 55.1% of their binding sites. This observation supports a model where CTCF and cohesin act in concert but can also act independently. Importantly, these data support a much greater CTCF-independent role for cohesin than previously reported.

4.2 Identification of the Canonical CTCF Motif

In order to confirm the specificity of our ChIP-seq, we sought to identify the known canonical CTCF binding motif [Schmidt et al., 2012], [Chen et al., 2008] in the sequence underlying the detected peaks. Identification of the motif was performed via the MEME-

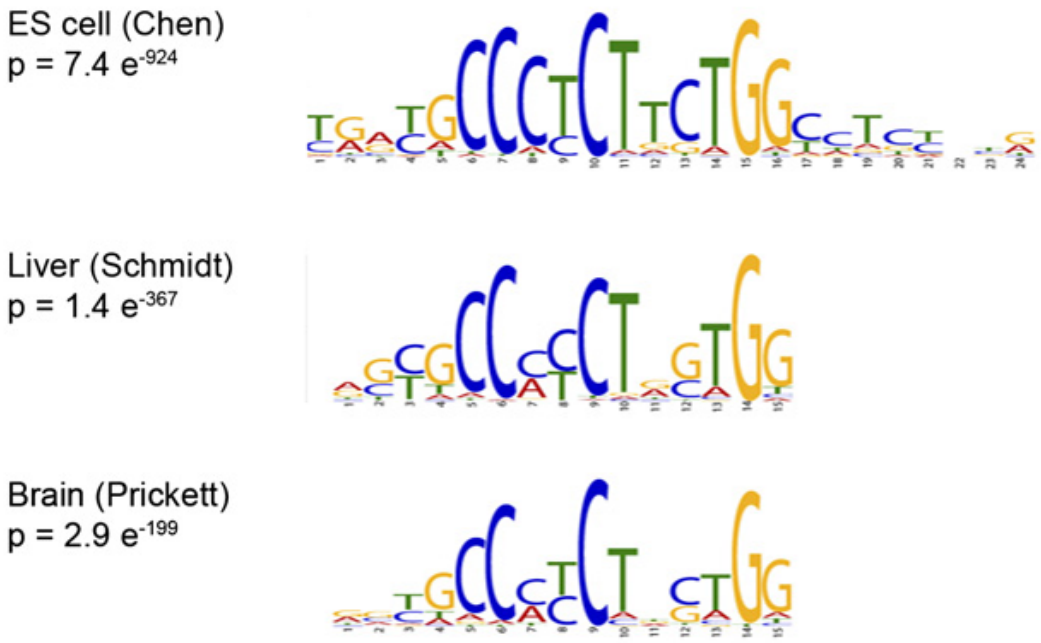


Figure 4.5: Results of CTCF motif discovery with MEME on ES Cells, liver and brain datasets. The canonical CTCF motif is identified in all three datasets with a high degree of confidence and shows high similarity between the three tissues. Adapted from [Prickett et al., 2013].

ChIP motif discovery tool [Bailey et al., 2009] on the brain dataset produced in the context of this study and datasets from liver and ES cells (Figure 4.5), as described in Section 2.8.11. The analysis revealed the canonical motif as the top hit (p-value=2.9e-199) in brain, and also confirmed its presence in other tissues. Our analysis did not identify the 12 bp second binding motif identified by Schmidt and colleagues [Schmidt et al., 2012].

Table 4.2: Cytosine methylation from [Xie et al., 2012] within CTCF binding sites identified in the present study, stratified by sequence context. CTCF binding sites are hypomethylated compared to the genome consistent with the known preference of CTCF for unmethylated DNA. Adapted from [Prickett et al., 2013].

	% Cytosine Methylation		
Cytosine context	Genome-wide	Within CTCF binding site	p-value
CpG	60.80%	51.90%	<1e-6
non-CpG	2.50%	2.10%	<1e-6

4.3 Examination of Methylation Status across CTCF Binding Sites

CTCF is known to preferentially bind unmethylated DNA and this specificity is responsible for its allele-specific activity at the *H19/Igf2* locus. We sought to confirm this relationship genome-wide.

Whole-genome methylation data for brain [Xie et al., 2012] were obtained and the overall level of methylation in the CpG and non-CpG context was compared to the level of methylation in CTCF binding regions (Table 4.2). As expected, cytosine methylation was significantly lower within CTCF binding peaks.

Unexpectedly, the difference in methylation in the non-CpG context was smaller than in the CpG context, suggesting that CpG methylation has greater influence on CTCF binding. This finding is somewhat surprising given that the canonical CTCF binding site does not contain CpGs. This difference can potentially be reconciled by the observation that the relative change in percent methylation is in both cases 0.15 of the genome-wide methylation, but the difference is exacerbated by the baseline difference in methylation in the two sequence contexts.

4.4 Tissue-specificity of CTCF Peaks without the Canonical Motif

CTCF has been reported as a multivalent DNA binding factor that can bind a diverse set of sequences using a different combinations of Zn-fingers [Filippova et al., 1996]. CTCF is however known to primarily bind a fixed motif, a finding replicated in the context of this study, and furthermore shown to be tissue invariant (Section 4.2), yet some CTCF binding locations do not contain the canonical motif sequence.

We investigated the relationship between tissue specificity and inclusion of the canonical motif in three tissues, using three datasets from liver [Schmidt et al., 2012], ES cells [Chen et al., 2008] and brain.

Initially, the number of overlapping peaks between the peak datasets was obtained. Given that the datasets were derived using different computational approaches and by different research groups we sought to find a peak size that will maximise specific overlaps without introducing excessive spurious hits. Optimisation of the peak size was performed by reducing the peak size in all samples to 1 bp and iteratively increasing it while calculating the number of overlaps observed. This optimisation was performed on an alternative ESC dataset from [Kagey et al., 2010], which was initially used in this project. The optimisation was not repeated in the dataset from Chen and colleagues [Chen et al., 2008] and is expected to be relevant as the inflection point was invariant for all the original datasets compared, suggesting that it would be similar for the dataset from Chen and colleagues.

The relationship between the number of overlapping peaks and the peak size is shown in Figure 4.6. The figure shows a rapid increase in the number of peak overlaps as the peak sizes increase from 0 to +/- 1kb and a plateau of non-specific overlaps after that point. For comparison, the same analysis was performed with our dataset (brain) against a randomly generated dataset to demonstrate that no inflexion point is present when interactions are random. On the basis of these results, a peak size of +/- 500bp was selected as this was the minimum size that coincided with the inflection point in all tissue combinations.

Using the size-adjusted peaks from all three datasets, we identified the number of overlaps between the three datasets (Figure 4.7 A) using the `intersectBed` tool. The three tissues share 25,269 peaks, which represents between 41.1% and 68.6% of the individual datasets. This level of overlap is consistent with an invariant role for CTCF in all cell types. Tissue-specific peaks for ES cells were considerably lower than for other tissues (5.1% against 29.1% in brain and 31.2% in liver), consistent with a basic state of differentiation and addition of new CTCF binding sites during specification.

In accordance with past observations, we hypothesised that the consensus CTCF motif is involved in tissue-invariant CTCF binding. Peaks containing the canonical motif were removed from each dataset to a cutoff of 10^{-4} using a custom pipeline utilising the FIMO motif matching tool (part of the MEME suite [Bailey et al., 2009]) as described in Section

2.8.11.

The results of the overlap analysis are shown in Figure 4.7 B. In sharp contrast to the previous analysis the number of overlapping peaks between the datasets is between 1.7% and 2.1%, whereas tissue specific peaks constitute the overwhelming majority of peaks. These results suggest that tissue-specific binding of CTCF may be dependent on tissue-specific motifs or recruitment of CTCF by other factors.

4.5 Identification of Tissue-specific CTCF Motifs

The observation that a very low number of peaks without the canonical motif are present in all three tissues examined raises the possibility that CTCF binds to tissue specific motifs, possibly synergistically with other factors. Such a suggestion would not be without precedent. In the original characterisation of CTCF, Fillipova and colleagues [Filippova et al., 1996] mutated specific Zn-fingers and demonstrated that CTCF can bind to heterologous sequences using different combinations of these DNA binding domains.

In order to investigate this possibility we performed *de novo* motif discovery analysis using the MEME-ChIP tool on tissue specific peaks [Bailey et al., 2009]. In addition to the canonical CTCF motif, this analysis identified a secondary motif in liver and ES cells with the sequence ACTCCAGTTCCAGGG with a high degree of confidence (Figure 4.8). A motif resembling this motif was also found in brain (Figure 4.8). The identification of a motif highly significant in three different independent datasets suggests that this motif may be biologically relevant. This motif does not resemble the secondary CTCF binding motif identified by Schmidt and colleagues, downstream of the canonical CTCF motif [Schmidt et al., 2012]. Furthermore, an additional sequence motif was found in ES cells (Figure 4.9) also at a high degree of confidence.

These findings indicate that secondary CTCF binding motifs may be present in tissue specific CTCF peaks that otherwise remain masked in analyses by the high abundance of the canonical motif. Further investigation in a more expansive set of tissue specific peaks is required to confirm these findings.

4.6 Analysis of Genome-wide Allele-specific Binding of CTCF

As aforementioned, the data generated in the context of this project was from animals derived from reciprocal crosses between Bl6 and *castaneous* mice strains. This allows for

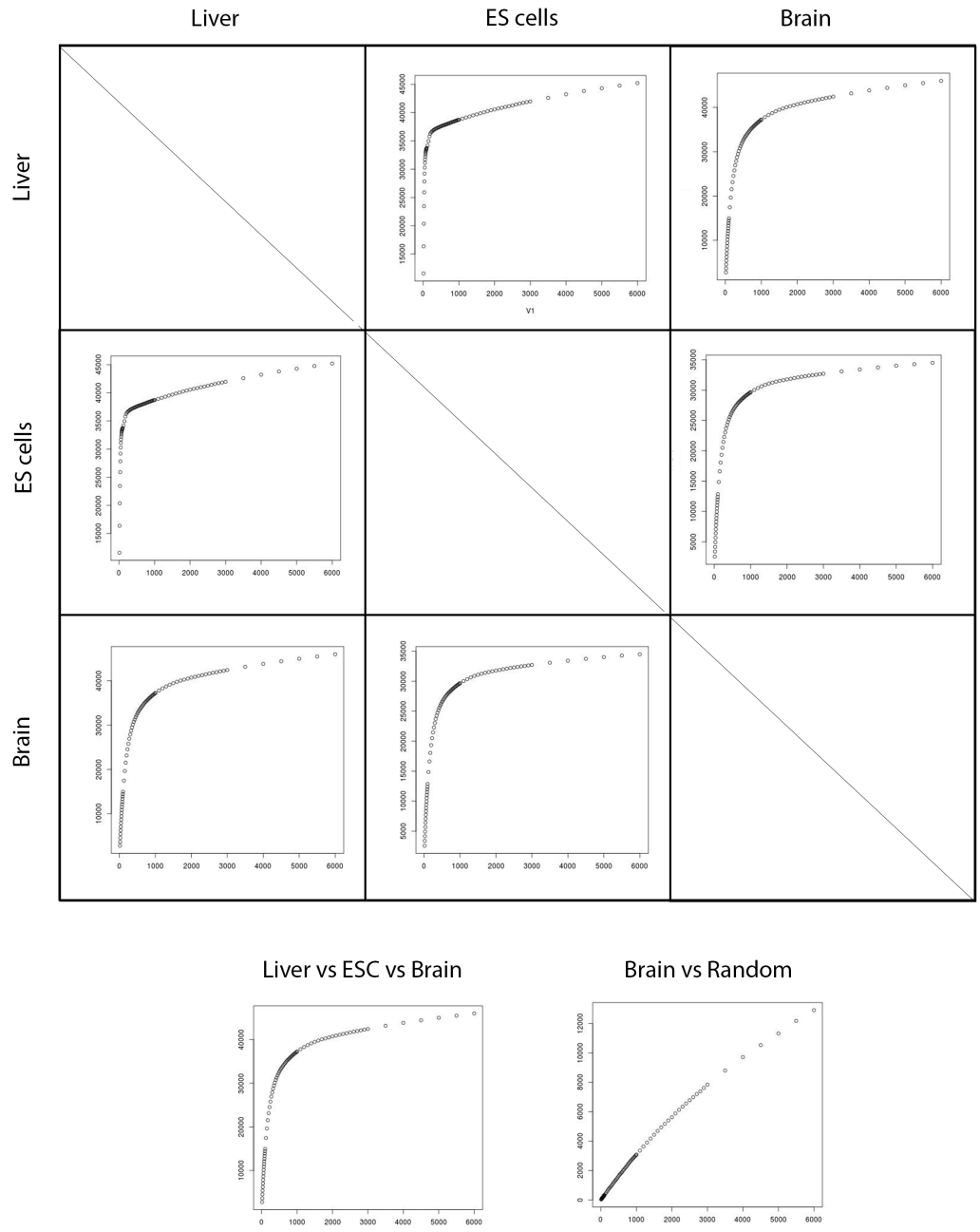


Figure 4.6: Plot of number of overlapping peaks against peak size for different dataset combinations suggests peak overlaps with a peak size smaller than 1 kb are specific and are not found in the comparison to a random dataset. In contrast, increase in overlaps beyond 1 kb appear to be largely random. On the basis of these plots a conservative peak size of 1 kb (+/- 500 bp) was used for the analysis.

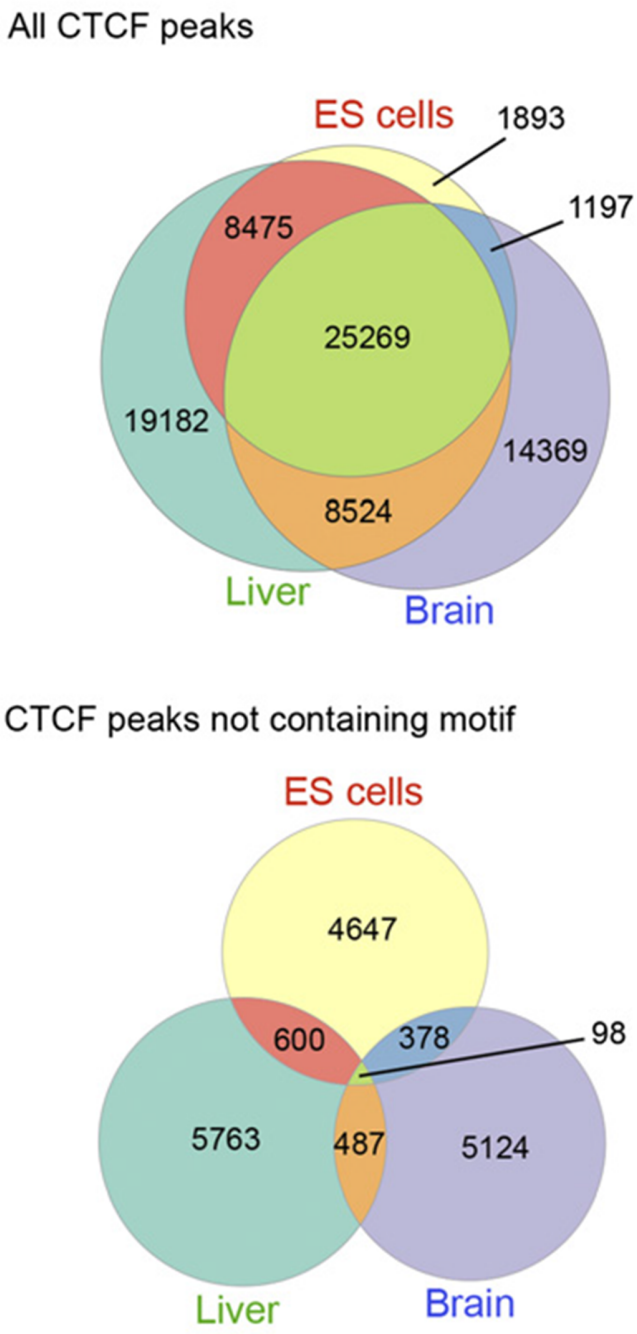


Figure 4.7: (A) Venn diagram of overlap of all CTCF peaks between ESC, liver and brain tissues. More than half of binding sites are shared between tissues suggesting a conserved function across tissues, after size adjustment. ESCs show the smallest number of unique CTCF peaks, consistent with an undifferentiated state. (B) Venn diagram of overlap of CTCF peaks not containing the canonical CTCF motif, shows poor overlap between tissues, suggesting binding to non-canonical motif is highly tissue specific. Adapted from [Prickett et al., 2013].

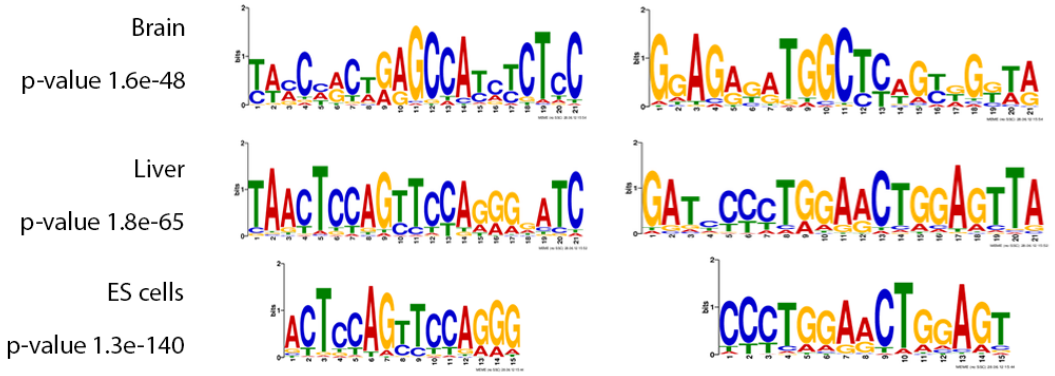


Figure 4.8: Motifs identified in tissue-specific CTCF peak sets. The motifs in ES cells and liver closely resemble each other.

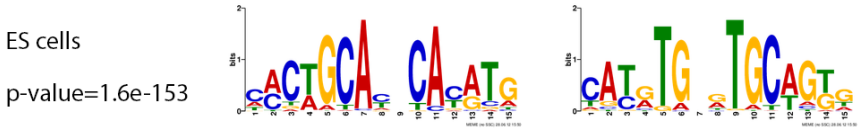


Figure 4.9: Motif identified in ES cell specific subset of CTCF peaks.

identification of mono-allelic binding events of CTCF and cohesin.

Individual reads overlapping CTCF or cohesin binding sites were examined for overlaps with known SNPs between the two parental mouse strains obtained from the Wellcome Trust Sanger Sequencing Centre Website [Yalcin et al., 2012]. Individual reads were assigned to one of the parental alleles on the basis of the best quality SNP (see following Section for details). Assignments from forward and reverse reads in read pairs were subsequently merged and the assignments to parental strains (Bl6 or Cast) were converted to assignments to parental origin (maternal or paternal) and data from reciprocal crosses were merged.

The counts of maternal and paternal assignments were generated for each CTCF or cohesin binding site. Allele-specific binding was assessed by means of a binomial test corrected for multiple testing via Bonferroni correction.

4.6.1 Read Assignment to Alleles

Assignment of reads to alleles was performed using a custom script (Appendix C.1). The approach is diagrammatically outlined in Figure 4.10. The known SNPs between the two parental strains were initially loaded into memory in a hash of arrays structure and sorted

for quick access (Figure 4.10, Step 1).

All reads were then processed individually and a binary search was performed for the start of each in the array corresponding to the aligned chromosome. The end of the read is then detected by linear search from the start position. Binary search for the end of the read was empirically found to be slower than linear search. This is because the number of SNPs in every read is small and also smaller than the number of steps required for the tracking of the end position in binary search.

After all the SNPs overlapping the read in question were identified, the SNP with the best read quality was sought and the assignment was performed on the basis of that polymorphism. In some cases assignment was not possible and the read was marked as unassigned.

4.6.2 Peak Size Adjustment and Filtering

In order to improve detection power CTCF and cohesin peaks were expanded by +/-500 bp and any overlaps arising were merged. This operation increased power to detect weak mono-allelic sites in close proximity by combining the peaks, but is expected to have negative impact in the detection of CTCF peaks of opposite parental allele binding in close proximity. Such a locus is not known to exist and this was not considered to be a significant caveat to the analysis. Expansion of peaks by 500 bp also has the advantage of ensuring that subsequent processing will take into account the assignment of reads partially overlapping the initially identified peak boundaries. The number of peaks before and after the expansion and merging of overlaps is shown in Figure 4.11.

4.6.3 Pipeline Optimisation

The assignment of individual reads to alleles is computationally intensive as it requires binary search of a large array of SNPs for every individual read. For this reason, the reads processed were filtered before assignment so as to exclude reads that do not overlap regions of CTCF or cohesin binding and to exclude reads that cannot be assigned due to a lack of known SNP between the parental strains. For similar reasons the number of peaks that did not contain a SNP between the parental strains were removed (Figure 4.11).

Only 47 million of the total 465 million reads sequenced overlaid peaks that could be assigned. Furthermore, only 38% of the reads in these regions overlaid a SNP and an

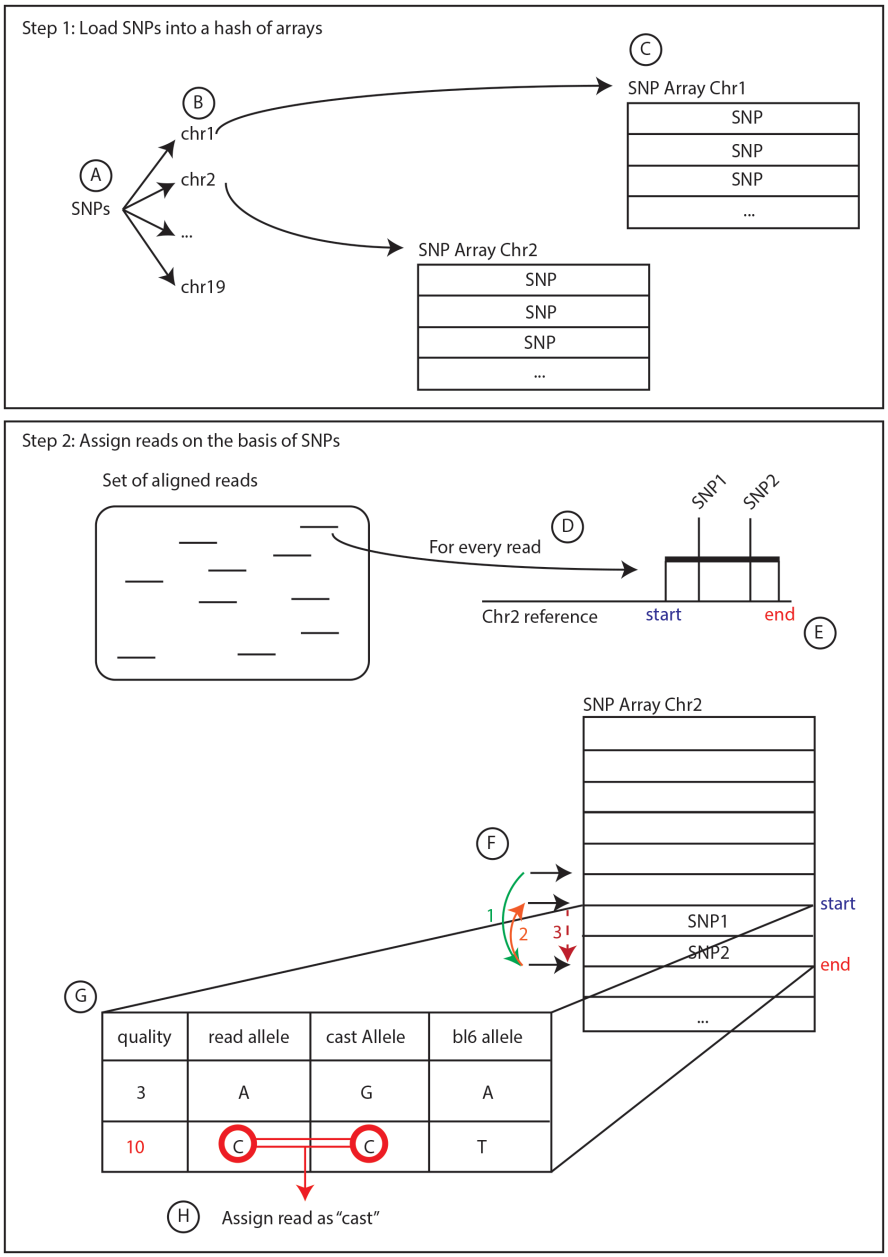


Figure 4.10: Schematic of algorithm for assignment of reads to alleles. The assignment is performed in two discrete steps. In step 1, the SNP information between the two parental strains is loaded onto memory and in step 2 individual reads are examined and assigned to parental alleles. In step 1 the SNP information is loaded into memory and saved in a hash (A) of per chromosome arrays (B) each containing SNP information sorted by chromosome position. In step 2, the start and end genomic position of every read (D) is retrieved and a binary search for it is performed against the SNP array for the relevant chromosome (F 1 and 2). The end position is found by linear search from the start position (F3). All the SNPs found between the start and end position are loaded into a temporary array (G) along with quality information from the read examined. The best quality position is selected and used to assign the read to a parental strain (H).

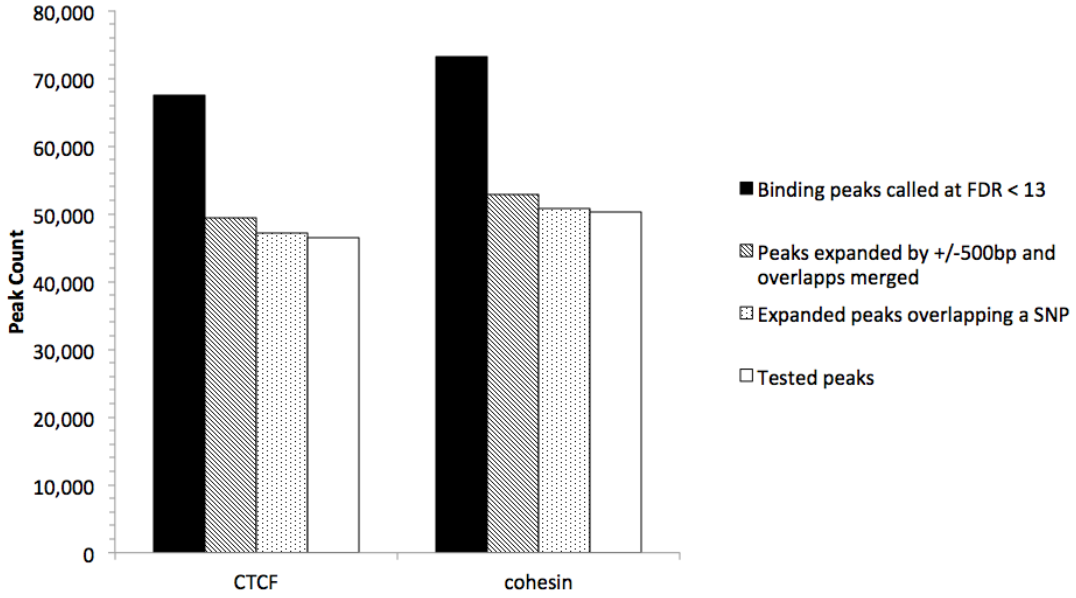


Figure 4.11: Count of processed CTCF and cohesin peaks across the analysis. A significant drop of the count of peaks occurs upon expansion and merging of peaks, but relatively few peaks do not overlap a SNP or an informative read. Adapted from [Prickett et al., 2013].

even smaller number (approximately 12 million) could be conclusively assigned (Figure 4.12). This observation exemplifies the importance of deep sequencing performed here for mono-allelic identification of peaks.

4.6.4 Reference Mapping Bias is Ameliorated by the Reciprocal Cross Experimental Design

A read mapping bias toward the reference (Bl6) allele was detected in the analysis (Figure 4.13 A) as expected. This bias is present because reads from the *castaneous* allele are less likely than reads from the Bl6 allele to align to the reference (Bl6) genome and an imbalance of aligned reads is generated. By employing a reciprocal cross design this bias was removed when read mapping was assigned to parental alleles and no overall bias remained (Figure 4.13 B).

4.6.5 Genome-wide Allele Specific CTCF and cohesin Binding Sites

The allele-specific analysis of binding sites of CTCF and cohesin described above, identified 21 genome-wide significant sites of parent of origin specific binding, shown in Table 4.3. Analysis of the cohesin binding sites did not identify any significant allele-specific

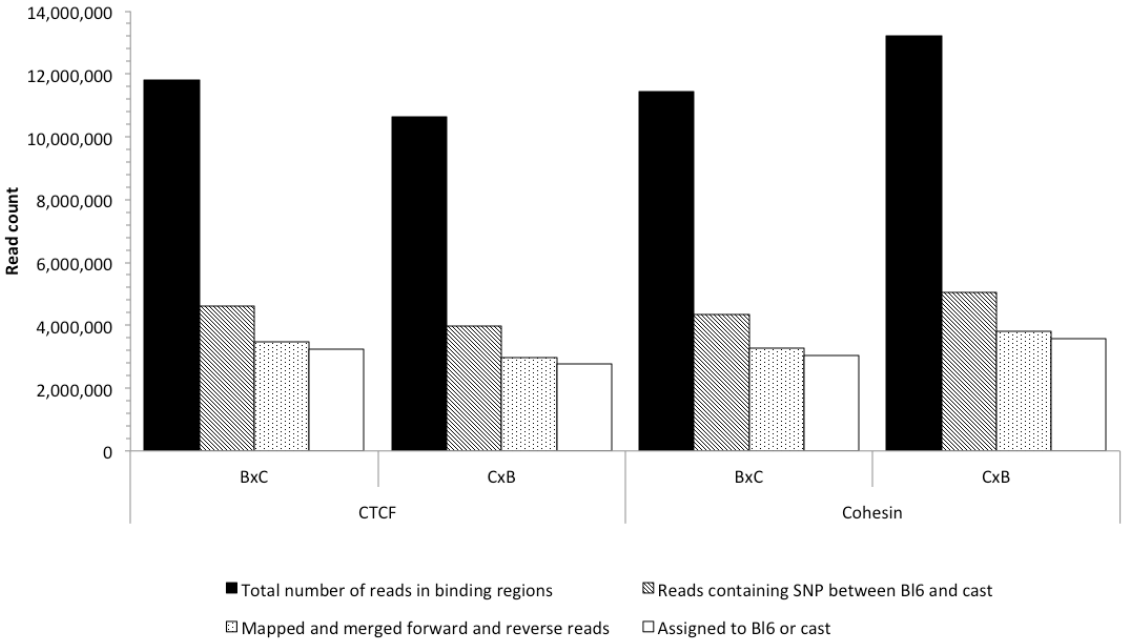


Figure 4.12: Read counts analysed across pipeline. Only a small portion of the total reads (not shown) overlays CTCF or cohesin peaks. The next single largest loss of reads occurs because more than half the reads cannot be assigned to a parental strain due to lack of a good quality informative SNP. Adapted from [Prickett et al., 2013].

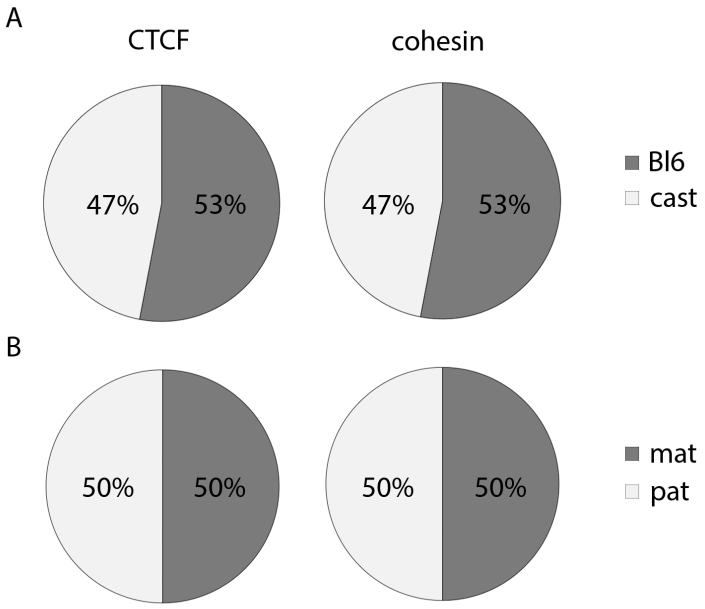


Figure 4.13: (A) Read assignments of CTCF and cohesin reads to parental strains. As expected, read assignment shows reference bias towards Bl6 allele. (B) Read assignments of CTCF and cohesin reads to parental origin. No residual parental origin bias remains after reciprocal cross data are merged. Adapted from [Prickett et al., 2013].

Table 4.3: Genome-wide significant regions of allele-specific CTCF binding. Adapted from [Prickett et al., 2013].

Genomic Locus	Binding Allele	Binding:Nonbinding	P-value	Nearest genes	Notes
chr7:149,764,416-149,768,874	maternal	10.26	1.11E-74	H19, Igf2	Known imprinted gene region
chr15:72,638,890-72,641,957	paternal	12.64	1.35E-57	Peg13, Trappc9	Known imprinted gene region
chr7:6,678,325-6,681,689	paternal	5.51	1.17E-30	Zim2 (Peg3)	Known imprinted gene region
chr6:30,686,300-30,688,046	paternal	9.38	9.10E-15	Mest	Known imprinted gene region
chr7:69,543,049-69,547,037	paternal	3.37	1.35E-12	Magel2	Known imprinted gene region
chr7:69,580,613-69,582,990	paternal	5.58	2.05E-10	Magel2	Known imprinted gene region
chr7:69,323,407-69,325,218	paternal	6.56	3.91E-10	Magel2	Known imprinted gene region
chr10:74,395,653-74,404,537	maternal	1.48	1.28E-09	Rtdr1, Gnaz	No known imprinted genes within 20Mb
chr7:69,526,343-69,528,366	paternal	4.11	3.10E-09	Magel2	Known imprinted gene region
chr7:69,372,124-69,373,922	paternal	8.00	3.26E-09	Ndn,Magel2	Known imprinted gene region
chr2:180,079,574-180,091,367	maternal	1.46	5.84E-09	Gata5, Gm14318	6Mb from Gnas locus
chr7:69,608,918-69,610,897	paternal	5.60	6.90E-09	Peg12, Mkrn3	Known imprinted gene region
chr14:69,941,084-69,946,555	paternal	1.52	1.68E-08	Gm16677, Entpd4, Loxl2	5Mb from Htr2a imprinted locus
chr7:69,353,580-69,355,185	paternal	5.09	2.15E-08	Ndn, Magel2	Known imprinted gene region
chr7:69,519,941-69,521,489	paternal	5.30	3.38E-08	Magel2	Known imprinted gene region
chr14:69,994,239-70,003,685	paternal	1.50	3.98E-08	Entpd4, AK086749, Loxl2	5Mb from Htr2a imprinted locus
chr10:120,737,183-120,739,873	paternal	2.73	4.75E-08	Tbc1d30	No known imprinted genes within 20Mb
chr3:121,236,161-121,244,419	maternal	1.67	6.85E-08	A530020G20Rik, Slc44a3	No known imprinted geneson chromosome 3
chr15:27,817,069-27,819,622	maternal	2.30	6.95E-08	Trio	No known imprinted genes within 20Mb
chr13:25,098,042-25,100,314	maternal	2.13	9.98E-08	Mrs2, Gpld1	No known imprinted genes within 20Mb
chr6:60,631,333-60,634,328	paternal	2.16	2.47E-07	Snca	1.6Mb from Herc3

Table 4.4: Top 20 monoallelic hits of cohesin binding. None of the hits shown here are significant after multiple testing correction. Only one of the hits coincides with a known imprinted locus (*H19*).

Genomic locus	Binding Allele	binding:nonbinding	p-value	Nearest Genes
chr7:134005379-134008113	paternal	3.9	3.85E-05	Hirip3
chr4:137237242-137240531	maternal	2.0	4.83E-05	Rap1gap
chr4:102979455-102980951	paternal	8.0	4.92E-05	Oma1
chr14:124619379-124623521	paternal	1.9	7.67E-05	Fgf14
chr5:112854608-112863824	paternal	1.4	7.99E-05	Sez6l
chr9:57939590-57940910	paternal	3.5	9.02E-05	Ccdc33
chr1:92481547-92493579	maternal	1.3	1.03E-04	Cops8
chr7:149763790-149770322	maternal	1.3	1.08E-04	H19
chr19:44445191-44450206	paternal	1.5	1.45E-04	Scd4
chr4:135878479-135881216	maternal	2.5	2.17E-04	Hnrnpr
chr5:148229197-148231785	maternal	1.8	2.21E-04	D5Ertd605e
chr6:42297399-42299927	maternal	3.5	2.47E-04	Zyx
chr9:103896310-103901797	paternal	1.4	2.62E-04	Nphp3
chr15:73789131-73790376	maternal	3.7	2.72E-04	Mroh5
chr1:170149664-170152248	paternal	2.4	3.17E-04	Pbx1
chr7:141756647-141760385	maternal	1.6	4.07E-04	Dock1
chr12:60064197-60066451	maternal	2.0	4.99E-04	Sec23a
chr1:62842761-62844278	paternal	5.5	5.34E-04	Nrp2
chr14:56146298-56147522	paternal	4.3	5.35E-04	Nrl
chr13:43186238-43191856	paternal	1.5	5.41E-04	Phactr1

binding sites, although several regions fell just short of the 1e-6 cutoff (Table 4.4).

Four of the known ICRs (*H19/Igf2*, *Peg13*, *Zim2* and *Mest*) were identified in the CTCF allele-specific analysis, demonstrating that our approach can detect known sites. Four other regions were within 6 Mb of loci known to contain imprinted genes (*Gnas*, *Htr2a* and *Herc3*). Eight of the other regions clustered near the imprinted *Peg12/Magel2* locus and the remaining five regions were not in the vicinity of known imprinted loci.

4.6.6 CTCF and cohesin Binding at Known DMRs

We interrogated CTCF and cohesin binding in the vicinity of 22 known DMRs. The results are shown in Table 4.5. We observed CTCF and/or cohesin binding in 19 of the 22 DMRs examined. Twelve sites were found to be bound by both proteins, three exclusively by CTCF and four exclusively by cohesin. No binding of either factor was found at three loci.

Mest and *Zim2* gDMRs were not previously known to bind CTCF in a parent-of-origin specific manner.

Given the very small portion of reads that were used for allele-specific assignment of peaks (Section 4.6.3), we considered the possibility that allele-specific binding occurs in some sites that did not exceed the multiple testing correction cutoff. The 95% confidence intervals of the read ratios were examined (Figure 4.14), this work was performed in collaboration with Dr Adam Prickett and Dr Jeniffer Mollon. A trend towards the binding to the unmethylated allele was observed for *Grb10*, *Mcts2*, *Cdh15*, *Nespas*, *Zrsr1*, *Peg10* and *Meg3/Dlk1* DMRs. Biallellic binding was observed at the *Inpp5f_v2* and *Plagl1* loci.

As aforementioned, no cohesin genome-wide significant binding sites were observed. Confidence interval analysis suggested that cohesin also tends to bind to the unmethylated allele in concert with CTCF, but exhibits an attenuated bias. This is consistent with a model whereby CTCF binds DNA in an allele-specific manner and subsequently recruits cohesin in a stochastic or temporally regulated manner.

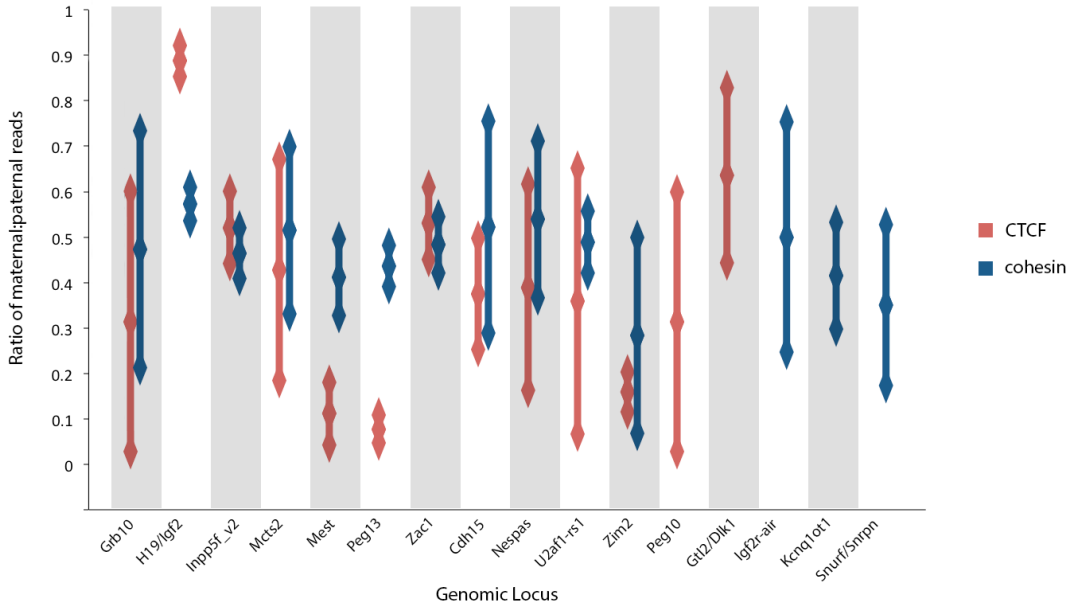


Figure 4.14: Plot of 95% confidence interval of ratio of maternal to paternal reads for CTCF and cohesin in the known imprinted loci shown in Table 4.5.

Table 4.5: CTFC and cohesin allele-specific binding in the vicinity of known DMRs. Adapted from [Prickett et al., 2013].

			CTCF and cohesin binding					
	gDMR information (Wamidex)		CTCF			Cohesin		
gDMR name	gDMR position	Methylated Allele	CTCF binding?	Binding allele	Allele specific p-value	Cohesin binding?	Binding allle	Allele binding p-value
CTCF and cohesin precisely colocalised at gDMR								
Grb10	chr11:11,925,485-11,925,790	Maternal	Yes	N/A	N/A	Yes	N/A	N/A
H19/Igf2	chr7:149,766,168-149,768,424	Paternal	Yes	Maternal*	1.11E-74*	Yes	Maternal	1.08E-04
Inpp5f_v2	chr7:135,831,788-135,832,156	Maternal	Yes	Bi-allelic	N/A	Yes	Bi-allelic	N/A
Mcts2	chr2:152,512,491-152,513,011	Maternal	Yes	N/A	N/A	Yes	N/A	N/A
Mest	chr6:30,686,709-30,687,273	Maternal	Yes	Paternal*	9.00E-15*	Yes	Paternal	0.0414
Nnat	chr2:157,385,786-157,387,398	Maternal	Yes	N/A	N/A	Yes	N/A	N/A
Peg13	chr15:72,636,765-72,642,079	Maternal	Yes	Paternal*	1.35E-57*	Yes	Paternal	5.50E-03
Plagl1	chr10:12,810,276-12,810,604	Maternal	Yes	Bi-allelic	N/A	Yes	Bi-allelic	N/A
CTCF and cohesin not precisely colocalised at gDMR								
Cdh15	chr8:125,387,861-125,390,344	Maternal	Yes	Paternal	0.0463	Yes	N/A	N/A
Nespas	chr2:174,121,208-174,126,482	Maternal	Yes	N/A	N/A	Yes	N/A	N/A
Zrsr1	chr11:22,871,842-22,872,319	Maternal	Yes	N/A	N/A	Yes	Bi-allelic	N/A
Zim2 (Peg3)	chr7:22,871,842-22,872,319	Maternal	Yes	Paternal*	1.16E-30*	Yes	Paternal	0.049
CTCF binding only								
Peg10	chr6:4,697,209-4,697,507	Maternal	Yes	N/A	N/A	No	N/A	N/A
Meg3/Dlk1	chr12:110,761,563-110,768,989	Paternal	Yes	N/A	N/A	No	N/A	N/A
Impact	chr18:13,130,706-13,132,250	Maternal	Yes	N/A	N/A	No	N/A	N/A
Cohesin binding only								
Igf2r-air	chr17:12,934,163-12,935,573	Maternal	No	N/A	N/A	Yes	N/A	N/A
Gnas-exon1A	chr2:174,153,279-174,153,502	Maternal	No	N/A	N/A	Yes	N/A	N/A
Kcnq1ot1	chr7:150,481,060-150,481,397	Maternal	No	N/A	N/A	Yes	N/A	N/A
Snurf/Snrpn	chr7:67,149,878-67,150,301	Maternal	No	N/A	N/A	Yes	N/A	N/A
No binding								
Nap1l5	chr6:58,856,690-58,857,056	Maternal	No	N/A	N/A	No	N/A	N/A
Rasgrf1	chr9:89,774,406-89,774,691	Paternal	No	N/A	N/A	No	N/A	N/A
Slc38a4	chr15:96,885,270-96,886,284	Maternal	No	N/A	N/A	No	N/A	N/A

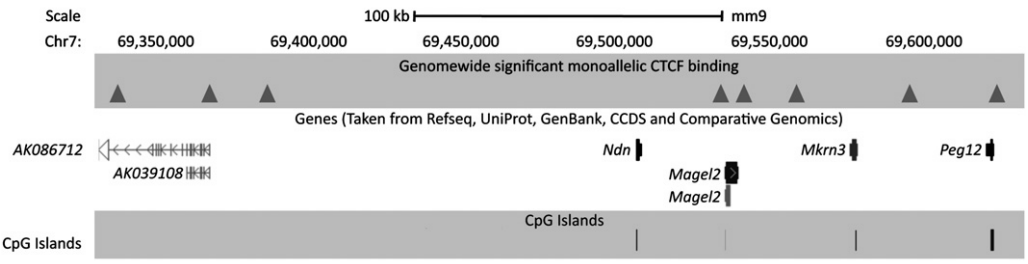


Figure 4.15: Paternal binding sites of CTCF binding near the *Magel2* locus (triangles). CpG islands are shown in the bottom track. This locus was the only locus identified by our analysis with eight monoallelic CTCF binding sites in close proximity. Reproduced from [Prickett et al., 2013].

4.6.7 The *Magel2* Locus

A particularly interesting result of this investigation was the observation that 8 CTCF paternal allele-specific binding sites cluster in the vicinity of the *Magel2* locus (Figure 4.15). Such organisation is not seen anywhere else in the genome in our analysis and has not been previously reported in the literature.

Given the above and that no known DMR existed in that locus at the time, the methylation of the CpG island at the promoter of *Magel2* was assayed via bisulphite conversion and colony PCR. This assay was performed by Ms Siohban Hughes as described in Section 2.8.12. The assay confirmed the presence of a maternally methylated DMR as shown in Figure 4.16. This DMR was independently reported in literature shortly after our analysis was complete [Xie et al., 2012].

4.7 Discussion

We performed high-depth allele-specific ChIP-seq of CTCF and the rad21 subunit of cohesin, two nuclear proteins that are known to bind in close proximity and have a known role in gene regulation, in the mouse brain at post-natal day 21. We searched for and identified the canonical CTCF motif in our peak set and we compared it with the canonical motif identified in liver and ES cells, confirming the specificity of our immunoprecipitation.

Comparison of the overlap of the binding sites of CTCF and rad21 showed considerable overlap of the binding sites of the two proteins, with over 50% of the sites between the two shared. This suggests that the two factors act in concert as has been established in the literature, but also have significant roles in isolation.

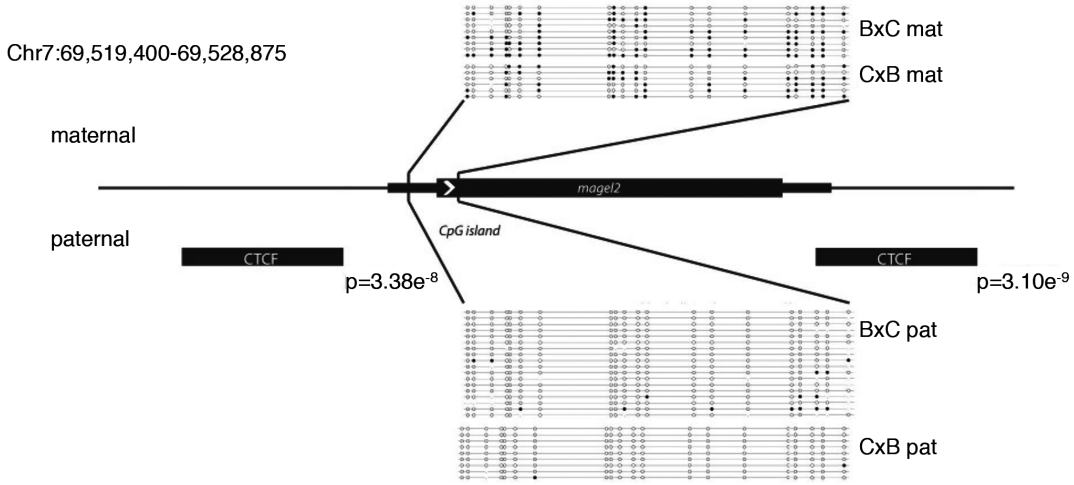


Figure 4.16: Methylation at the *Magel2* promoter CpG island shows parent-of-origin specific methylation of the maternal allele. Filled and empty cycles represent methylated and unmethylated CpGs respectively. Adapted from [Prickett et al., 2013].

CTCF is known to have a higher affinity for unmethylated DNA. Using brain methylation data from others [Xie et al., 2012] we compared the methylation over CTCF binding regions with methylation on non CTCF binding regions. Consistent with the affinity of CTCF for unmethylated DNA, regions found by our assay to bind CTCF, displayed lower methylation levels than the genomic mean. Unexpectedly the difference in methylation was more profound in the CpG context. This was despite that no CpG dinucleotides appear in the canonical CTCF motif. The discrepancy between the difference in methylation in CpG and non-CpG content may be attributable to the lower genome-wide methylation of non-CpG Cs.

We investigated the tissue specificity of CTCF peaks, by comparing our dataset with datasets from liver and ES cells. We found a significant overlap between different cell types, but also a significant proportion of tissue specific peaks. Tissue-specific peaks in ES cells were many fewer that in other tissues. This is consistent with a basal set of CTCF binding peaks in ES cells that is expanded and refined during differentiation.

Furthermore, we investigated the tissue specificity of CTCF peaks after subtraction of peaks containing the canonical CTCF motif. We observed a very different pattern than reported above. The overlap of peaks not containing the canonical motif between tissues was found to be very poor. This suggests that the canonical motif is implicated in binding in peaks that are tissue invariant, whereas tissue specific peaks are the result of a separate binding mechanism, such as binding to a different motif or recruitment by other factors. CTCF is known to be able to use different combination of Zn-fingers to

bind divergent sequences [Filippova et al., 1996].

This prompted us to perform *de novo* motif discovery on the tissue-specific CTCF peaks. We independently identified a novel motif in all three tissues with high degree of confidence. Furthermore, we identified a tissue specific motif in ES cells. The biological relevance of these motifs is unclear but their discovery warrants further biochemical investigation.

Making use of the allele-specific nature of our data, we developed a pipeline for the identification of mono-allelic binding events and we applied it to the CTCF and cohesin datasets. We identified 21 regions of parent-of-origin allele-specific binding for CTCF. The majority (16/21) of the identified allele-specific peaks overlaid or where in the vicinity of a known imprinted locus. Eight of the identified loci were in close proximity near the *Peg12/Magel2* locus in an unprecedented configuration. No allele-specific peaks were identified for cohesin.

We cross-referenced our allele-specific data with the list of known DMRs. We observed CTCF and/or cohesin binding in 19 of the 22 DMRs. Only about half the sites were found to be bound by both proteins and a smaller portion exclusively by one of the proteins. This finding supports the notion that binding is mechanistically diverse. The timepoint of establishment of DMRs was not a determinant of which factors were bound to it, suggesting that the mechanistic diversity independent of the establishment mechanism.

Allele-specific binding of CTCF was observed in five DMRs in a pattern consistent with the known methylation at these sites, with CTCF appearing bound to the unmethylated allele in all cases. Cohesin allele-specific binding was not observed. Confidence interval analysis demonstrated a more polarised binding pattern for CTCF than cohesin in the majority of the DMRs examined. This is consistent with a model where allele-specific binding of CTCF influences, but does not directly dictate cohesin binding.

4.7.1 Further Work

The work presented in this chapter presents further questions in both the context of specific loci and genome-wide.

The tissue-specific motif analysis has resulted in the identification of two previously unappreciated motifs. The biological relevance of these motifs is unknown and further investigation is in order. It would be of particular interest to assess the affinity of CTCF for the motif found by biochemical assays. Our analysis has given rise to the distinct

possibility that other tissue specific motifs are present in tissues other than the ones examined. A comprehensive identification of CTCF ChIP-seq data sets in other tissues in the literature and application of our analysis on these would thus be of interest.

CTCF monoallelic binding in sites with no known imprinted genes may point to novel imprinted genes. Validation of these loci was performed by others subsequent to the work presented here and did not result in the identification of any novel imprinted genes. Excluding the possibility that these loci represent false positives in our analysis this suggests that the loci in question may be evolutionary remnants of imprinted regions, that still maintain an allele-specific chromosomal organisation but do not display imprinted expression. Alternatively, these could be tissue-specific imprinted loci outside of the brain.

Finally, the *Magel2* locus identified in this study presents the identification of an unprecedented genomic organisation with eight paternal binding sites with 200 kb. This unusual genomic organisation warrants further investigation. The presence of this organisation patterns should be assessed in other tissues, followed by a comprehensive investigation into the transcripts resulting from this locus, accompanied by their allelic-specificity. Generation of allele-specific chromatin conformation data for this locus would also be of interest.

Chapter 5

Discussion

5.1 Overview

The modern definition of epigenetics encompasses a diverse range of biological phenomena that share the common feature of providing a mechanism of stable, and according to some definitions heritable across cell divisions, storage of information in cells other than that encoded in the primary DNA sequence.

The work presented in this thesis utilises next-generation sequencing and bioinformatics methodologies to investigate the interplay between epigenetic and transcriptional processes in two different biological systems. The systems examined are sufficiently diverse so as to inform on the expansive and diverse role of epigenetics, while addressing outstanding specific questions of interest in their particular areas.

Specifically, the methylome and transcriptome of the specialised heart cell line comprising the endocardium, are examined and are juxtaposed with those of other endothelial cells. Furthermore, the genomic distributions of CTCF and cohesin, proteins implicated in a diverse set of biological process including transcription and imprinting are investigated in an allele-specific manner and integrated with genome-wide methylation data from other independent studies in whole brain.

The work presented here is timely given the advent of next-generation sequencing technologies and the emergence of bioinformatics methodologies that for the first time allow the investigation of epigenetic marks and transcription in the whole genome scale at a relatively low cost. Other studies based on genome-wide interrogation of the transcriptome and epigenome have revealed methylation changes in the specification of embryonic stem cells to germ layers [Gifford et al., 2013] and methylome changes later in blood

differentiation [Zilbauer et al., 2013].

The work presented here expands on our understanding of the epigenetic landscape during development that has been established by a multitude of independent studies. Large concerted efforts such as the ENCODE and BluePrint projects are addressing similar questions in large scale.

5.2 Epigenetics and Transcriptomics in Endocardial and Endothelial Differentiation

We examined the transcriptional and epigenetic landscape of endocardial and endothelial cells. The endocardium is a specialised type of epithelium that lines the atria and ventricles of the developing and adult heart. It is a tissue of particular interest because it is involved in a number of processes of cardiac development including trabeculation of the ventricles, septation of the heart, valve leaflet and coronary circulation formation [Harris and Black, 2010] [Tian et al., 2014]. Given the contribution of this cell lineage to heart development and given that congenital abnormalities are amongst the most common causes of early morbidity and mortality, understanding the role of the endocardium is critical.

Our initial hypothesis was that epigenetic differences and in particular differential DNA methylation were primarily responsible for cell-fate decisions and specification of the endocardial lineage. We expected to observe only subtle differences in gene expression, consistent with epigenetics playing a role in tissue specific gene expression and differentiation. This hypothesis was based on the identical functionality and highly similar morphology of the endocardium and the endothelium in early but not late development that suggested the presence of latent information in these cells, prior data that did not identify any transcriptional differences between the two cell lines and several studies suggesting extensive contribution of epigenetic processes in cardiac development in general [Baccarelli et al., 2010] [Mathiyalagan et al., 2010] [Paige et al., 2012] [Wamstad et al., 2012] [Chang and Bruneau, 2012] [Vallaster et al., 2012].

We addressed this question using an *in vitro* differentiation model of the endocardium, that has been shown to recapitulate endocardial development [Narumiya et al., 2007] [Misfeldt et al., 2009], and contrasted this cell line to cultured endothelial cells from the same system, using a specific marker of the endocardium (NFATc1) to isolate these cells via flow cytometry.

In contrast to past data and our expectations, we observed a considerable number of both transcriptomic changes and methylome changes, suggesting that both gene regulation and epigenetic processes have a role in endocardial differentiation.

Using the generated transcriptomic dataset and bioinformatics methodologies, we identify specific members of the ETS family of transcription factors, a family that has been independently implicated in hematopoietic and vascular development, as likely to play an important role in the development of the endocardium. Furthermore, we prioritised the identified differentially regulated genes using different criteria for further experimental assessment of their functional significance.

We assessed methylome changes on a per CGI basis as well as genome-wide after using a smoothing algorithm to improve our power to detect differential methylation events. The genomic distribution of the observed methylome changes was assessed and an over-representation of differentially methylated CGIs in the vicinity of protein coding regions was observed. This was in agreement with past results by others, suggesting that terminal differentiation methylome changes occur primarily intragenically [Deaton et al., 2011]. On the genome-wide level we observed enrichment of DMRs in coding regions but also overlap with functional genomic elements. Furthermore, the genes that were in the vicinity of DMRs were significantly associated with GO Terms strongly suggesting developmental role and therefore providing evidence for the role of the DMRs in endocardial differentiation.

Furthermore, interplay between the methylome and transcriptional regulation was observed. We were able to confirm that the well-described pattern of inverse correlation between CGI methylation and transcription is observed in each tissue individually and examined the nature of this relationship as a function of genomic distance. This, in conjunction with the expected methylation at the vast majority of the known imprinted loci, reaffirms the known connections between the epigenome and gene regulation and supports the robustness and relevance of both the datasets generated here. We also examine overlaps between the set of genomic DMRs identified and the set of differentially expressed genes detected. These overlaps included genes that are well known regulators of the endocardium, such as *Tie1* and *Tal1*, potentially providing a link between epigenetic regulation and endocardial development.

The observations presented in the context of this work suffer from the caveats discussed extensively in Section 3.5.1. Importantly, the design of this study was such that it did not

allow the analysis at the time of specification. The endocardial cells were examined after differentiation occurred, at a time point at which initiating event in the endocardial specification may not persist. The methylation assay furthermore suffers from the limitation that hydroxylmethylation can not be distinguished from DNA methylation. Although hydroxylmethylation is not present in most differentiated tissues, it may be relevant in this context because the endocardial cells examined here are directly derived from differentiation of ES cells and hydroxylmethylation may persist. Finally, the data generated were not validated using locus specific assays or by functional downstream analysis. Ideally, validation of the results presented here would be performed *in vivo* endocardium to establish their validity in this context before functional validation in the form of knockdown of identified expressed factors and genetic manipulation of differentially methylated loci is pursued.

Despite the above, the work supports a role for both the methylome and the transcriptome in the differentiation of the endocardium and suggests that in the future the two should be examined in concert to obtain a complete understanding of the differentiation of this tissue. The results furthermore support the idea that other epigenetic marks are involved in this differentiation process and provide justification for further studies of the epigenetics of the endocardium.

5.3 Allele-specific CTCF and cohesin Binding in the Mouse Brain and the Role of DNA Methylation

We assessed the genome-wide distribution of CTCF and cohesin in an allele-specific manner by performing high-depth ChIP-seq for these two factors in post-natal day 21 whole brain from the offspring of Bl6 and *casteneous* mice strains. We related the generated data to gene expression in the genomic context of imprinted genes, the correct expression of which is critical for development [Surani et al., 1984] [McGrath and Solter, 1984]. Brain was utilised because a large portion of all known imprinted genes are expressed in this tissue [Schulz et al., 2008] [Wilkins, 2008].

CTCF is a Zn-finger protein traditionally categorised as an insulator, but with known roles in the regulation of gene expression, chromosomal organisation and regulation of imprinting [Phillips and Corces, 2009], most prominently at the *H19* locus [Kanduri et al., 2000]. CTCF is known to preferentially bind unmethylated DNA and can, through

selective DNA binding, link DNA methylation and gene regulation. The ubiquity of CTCF, its genome-wide organisational function and methylation sensitivity has led to the proposal that CTCF is part of a heritable epigenetic system [Phillips and Corces, 2009], a claim that has not so far been substantiated.

Cohesin is a protein complex with an established role in sister chromatid cohesion during nuclear division [Onn et al., 2008] [Ocampo-Hafalla and Uhlmann, 2011] that is known to associate closely with CTCF [Parelho et al., 2008] [Wendt et al., 2008] and is also believed to regulate gene expression.

In addition to the examination of the interplay between epigenetics in the form of DNA methylation and the binding of these transcription associated factors, we addressed questions that this system presents in its own right. Specifically, we assessed the homogeneity of imprinting mechanisms as judged by the presence of CTCF and cohesin and under the assumption that CTCF binds a significant portion of imprinted loci in an allele-specific manner, we searched for the existence of novel and previously uncharacterised imprinted loci.

In line with observations made in other studies [Wendt et al., 2008] [Lin et al., 2011], we observed that CTCF and cohesin share approximately half their binding sites, suggesting that the two factors can work synergistically but also in isolation. We identified the canonical CTCF motif in sites of CTCF binding, reaffirming the robustness of our data. Using publicly available data from other studies we confirmed the genome-wide preference for CTCF for unmethylated DNA [Xie et al., 2012]. Furthermore, we assessed the tissue specificity of CTCF peaks by comparison with ChIP-seq data from liver and brain [Schmidt et al., 2012] [Chen et al., 2008]. We observed significant overlap, but also noted a paucity of ES cell specific peaks suggesting that the binding repertoire of CTCF expands upon differentiation in a tissue specific manner. The peak overlap of CTCF in different tissues was found to be very poor when only CTCF peaks not containing the canonical motif were examined, pointing towards non-sequence specificity mechanisms from binding at these sites.

We utilised the allele-specific nature of our data to investigate the parent-of-origin specific binding of CTCF and cohesin in a genome-wide scale. After multiple testing correction, we observed 21 sites of allele-specific binding many of which were in the vicinity of known imprinted loci. Further examination of the *Magel2* locus revealed the presence of eight CTCF binding peaks, a configuration not encountered anywhere else in the genome.

We confirmed the presence of a novel DMR in the *Magel2* locus, which was shortly after independently reported in literature.

We examined the binding of CTCF and cohesin near known imprinted loci. We observed high heterogeneity in the binding of CTCF and cohesin. CTCF was found to bind to the unmethylated allele in all cases where parent-of-origin specific binding was observed, suggesting a strong link between epigenetics and imprinted expression in this system. Cohesin was not found to be bound allele-specifically to any genomic locations inclusive of imprinted loci, although its binding was biased toward the binding allele of CTCF in all cases, suggesting recruitment by CTCF.

Overall, our results point towards a heterogenous role of regulation of expression by CTCF and cohesin, which however displays well-defined mechanistic properties in allele-specific DNA binding.

5.4 Concluding Remarks

Two dissimilar tissues were examined in the context of this work. The methylome and transcriptome of the highly specialised endocardium in culture at the embryonic day equivalent of E9.5 and genome-wide CTCF and cohesin parent-of-origin specific binding in whole brain at post-natal day 21. The two systems contrast sharply in a number of features, namely developmental time-point, germ-layer, tissue specificity level and *in vitro* against *in vivo* setting. In both cases we observe interplay between the methylome and transcriptional processes, supporting the notion that examination of both processes is important for complete understanding of gene regulation and cellular behviour.

Bibliography

[Abu-Issa and Kirby, 2007] Abu-Issa, R. and Kirby, M. L. (2007). Heart field: From mesoderm to heart tube. *Annual Review of Cell and Developmental Biology*, 23(1):45–68.

[Anders and Huber, 2010] Anders, S. and Huber, W. (2010). Differential expression analysis for sequence count data. *Genome Biology*, 11(10):R106.

[Andrews, 2014] Andrews, S. R. (2014). FASTQC - a quality control tool for high throughput sequence data. *unpublished*. http://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

[Arand et al., 2012] Arand, J., Spieler, D., Karius, T., Branco, M. R., Meilinger, D., Meissner, A., Jenuwein, T., Xu, G., Leonhardt, H., Wolf, V., and Walter, J. (2012). *In vivo* control of CpG and non-CpG DNA methylation by DNA methyltransferases. *PLoS Genetics*, 8(6):e1002750.

[Ásgeirsdóttir et al., 2012] Ásgeirsdóttir, S. A., van Solingen, C., Kurniati, N. F., Zwiers, P. J., Heeringa, P., van Meurs, M., Satchell, S. C., Saleem, M. A., Mathieson, P. W., Banas, B., Kamps, J. A. A. M., Rabelink, T. J., van Zonneveld, A. J., and Molema, G. (2012). MicroRNA-126 contributes to renal microvascular heterogeneity of VCAM-1 protein expression in acute inflammation. *American Journal of Physiology - Renal Physiology*, 302(12):F1630–F1639.

[Baccarelli et al., 2010] Baccarelli, A., Rienstra, M., and Benjamin, E. J. (2010). Cardiovascular epigenetics: Basic concepts and results from animal and human studies. *Circulation: Cardiovascular Genetics*, 3(6):567–573.

[Bailey et al., 2009] Bailey, T. L., Boden, M., Buske, F. A., Frith, M., Grant, C. E., Clementi, L., Ren, J., Li, W. W., and Noble, W. S. (2009). MEME suite: tools for motif discovery and searching. *Nucleic Acids Research*, 37(suppl 2):W202–W208.

[Baldwin, 1996] Baldwin, H. (1996). Early embryonic vascular development. *Cardiovascular research*, 31 Spec No:E34–45.

[Baniahmad et al., 1990] Baniahmad, A., Steiner, C., Köhne, A. C., and Renkawitz, R. (1990). Modular structure of a chicken lysozyme silencer: Involvement of an unusual thyroid hormone receptor binding site. *Cell*, 61(3):505 – 514.

[Banjo et al., 2013] Banjo, T., Grajcarek, J., Yoshino, D., Osada, H., Miyasaka, K. Y., Kida, Y. S., Ueki, Y., Nagayama, K., Kawakami, K., Matsumoto, T., Sato, M., and Ogura, T. (2013). Haemodynamically dependent valvulogenesis of zebrafish heart is mediated by flow-dependent expression of miR-21. *Nature Communications*, 4(1978).

[Bannister and Kouzarides, 2011] Bannister, A. J. and Kouzarides, T. (2011). Regulation of chromatin by histone modifications. *Cell Research*, 21(3):381–395.

[Bäumer et al., 2006] Bäumer, S., Keller, L., Holtmann, A., Funke, R., August, B., Gamp, A., Wolburg, H., Wolburg-Buchholz, K., Deutsch, U., and Vestweber, D. (2006). Vascular endothelial cell-specific phosphotyrosine phosphatase (VE-PTP) activity is required for blood vessel development. *Blood*, 107(12):4754–4762.

[Behrens et al., 2014] Behrens, A. N., Zierold, C., Shi, X., Ren, Y., Koyano-Nakagawa, N., Garry, D. J., and Martin, C. M. (2014). *Sox7* is regulated by ETV2 during cardiovascular development. *Stem Cells and Development*, 23(17):2004–2013.

[Bell et al., 1999] Bell, A. C., West, A. G., and Felsenfeld, G. (1999). The protein CTCF is required for the enhancer blocking activity of vertebrate insulators. *Cell*, 98(3):387–396.

[Berger et al., 2009] Berger, S. L., Kouzarides, T., Shiekhattar, R., and Shilatifard, A. (2009). An operational definition of epigenetics. *Genes & Development*, 23(7):781–783.

[Bestor et al., 1988] Bestor, T., Laudano, A., Mattaliano, R., and Ingram, V. (1988). Cloning and sequencing of a cDNA encoding DNA methyltransferase of mouse cells: The carboxyl-terminal domain of the mammalian enzymes is related to bacterial restriction methyltransferases. *Journal of Molecular Biology*, 203(4):971 – 983.

[Bhutani et al., 2011] Bhutani, N., Burns, D. M., and Blau, H. M. (2011). DNA demethylation dynamics. *Cell*, 146(6):866 – 872.

[Bird, 1980] Bird, A. P. (1980). DNA methylation and the frequency of CpG in animal DNA. *Nucleic Acids Research*, 8(7):1499–1504.

[Blow et al., 2010] Blow, M. J., McCulley, D. J., Li, Z., Zhang, T., Akiyama, J. A., Holt, A., Plajzer-Frick, I., Shoukry, M., Wright, C., Chen, F., Afzal, V., Bristow, J., Ren, B., Black, B. L., Rubin, E. M., Visel, A., and Pennacchio, L. A. (2010). ChIP-Seq identification of weakly conserved heart enhancers. *Nature Genetics*, 42(9):806–810.

[Bock, 2012] Bock, C. (2012). Analysing and interpreting DNA methylation data. *Nature Review Genetics*, 13(10):705–719.

[Bogachek et al., 2014] Bogachek, M. V., De Andrade, J. P., and Weigel, R. J. (2014). Regulation of epithelial–mesenchymal transition through sumoylation of transcription factors. *Cancer Research*, 75(1):12–15.

[Bonachea et al., 2014] Bonachea, E. M., Chang, S.-W., Zender, G., LaHaye, S., Fitzgerald-Butt, S., McBride, K. L., and Garg, V. (2014). Rare GATA5 sequence variants identified in individuals with bicuspid aortic valve. *Pediatr Res*, 76(2):211–216.

[Booth et al., 2013] Booth, M. J., Ost, T. W. B., Beraldi, D., Bell, N. M., Branco, M. R., Reik, W., and Balasubramanian, S. (2013). Oxidative bisulfite sequencing of 5-methylcytosine and 5-hydroxymethylcytosine. *Nat. Protocols*, 8(10):1841–1851.

[Bostick et al., 2007] Bostick, M., Kim, J. K., Estève, P.-O., Clark, A., Pradhan, S., and Jacobsen, S. E. (2007). UHRF1 plays a role in maintaining DNA methylation in mammalian cells. *Science*, 317(5845):1760–1764.

[Burcin et al., 1997] Burcin, M., Arnold, R., Lutz, M., Kaiser, B., Runge, D., Lottspeich, F., Filippova, G. N., Lobanenkov, V. V., and Renkawitz, R. (1997). Negative protein 1, which is required for function of the chicken lysozyme gene silencer in conjunction with hormone receptors, is identical to the multivalent zinc finger repressor CTCF. *Molecular and Cellular Biology*, 17(3):1281–8.

[Bussmann et al., 2007] Bussmann, J., Bakkers, J., and Schulte-Merker, S. (2007). Early endocardial morphogenesis requires Scl/Tal1. *PLoS Genetics*, 3(8):e140.

[Cai et al., 2003] Cai, C.-L., Liang, X., Shi, Y., Chu, P.-H., Pfaff, S. L., Chen, J., and Evans, S. (2003). Isl1 identifies a cardiac progenitor population that proliferates prior

to differentiation and contributes a majority of cells to the heart. *Developmental Cell*, 5(6):877–889.

[Chakraborty et al., 2010] Chakraborty, S., Combs, M., and Yutzey, K. (2010). Transcriptional regulation of heart valve progenitor cells. *Pediatric Cardiology*, 31(3):414–421.

[Chambers and Tomlinson, 2009] Chambers, I. and Tomlinson, S. R. (2009). The transcriptional foundation of pluripotency. *Development*, 136(14):2311–2322.

[Chan et al., 2004] Chan, Y., Fish, J. E., D'Abreo, C., Lin, S., Robb, G. B., Teichert, A.-M., Karantzoulis-Fegaras, F., Keightley, A., Steer, B. M., and Marsden, P. A. (2004). The cell-specific expression of endothelial nitric-oxide synthase: A role for dna methylation. *Journal of Biological Chemistry*, 279(33):35087–35100.

[Chang and Bruneau, 2012] Chang, C.-P. and Bruneau, B. G. (2012). Epigenetics and cardiovascular development. *Annual Review of Physiology*, 74(1):41–68.

[Chen et al., 2013] Chen, H., Zhang, W., Sun, X., Yoshimoto, M., Chen, Z., Zhu, W., Liu, J., Shen, Y., Yong, W., Li, D., Zhang, J., Lin, Y., Li, B., VanDusen, N. J., Snider, P., Schwartz, R. J., Conway, S. J., Field, L. J., Yoder, M. C., Firulli, A. B., Carlesso, N., Towbin, J. A., and Shou, W. (2013). Fkbp1a controls ventricular myocardium trabeculation and compaction by regulating endocardial notch1 activity. *Development*, 140(9):1946–1957.

[Chen et al., 2008] Chen, X., Xu, H., Yuan, P., Fang, F., Huss, M., Vega, V. B., Wong, E., Orlov, Y. L., Zhang, W., Jiang, J., Loh, Y.-H., Yeo, H. C., Yeo, Z. X., Narang, V., Govindarajan, K. R., Leong, B., Shahab, A., Ruan, Y., Bourque, G., Sung, W.-K., Clarke, N. D., Wei, C.-L., and Ng, H.-H. (2008). Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. *Cell*, 133(6):1106 – 1117.

[Chernukhin et al., 2007] Chernukhin, I., Shamsuddin, S., Kang, S. Y., Bergström, R., Kwon, Y.-W., Yu, W., Whitehead, J., Mukhopadhyay, R., Docquier, F., Farrar, D., Morrison, I., Vigneron, M., Wu, S.-Y., Chiang, C.-M., Loukinov, D., Lobanenkov, V., Ohlsson, R., and Klenova, E. (2007). Ctcf interacts with and recruits the largest subunit of rna polymerase ii to ctcf target sites genome-wide. *Molecular and Cellular Biology*, 27(5):1631–1648.

[Chiavegatto et al., 2012] Chiavegatto, S., Sauce, B., Ambar, G., Cheverud, J. M., and Peripato, A. C. (2012). Hypothalamic expression of Peg3 gene is associated with maternal care differences between SM/J and LG/J mouse strains. *Brain and Behavior*, 2(4):365–376.

[Chim et al., 2014] Chim, S. M., Kuek, V., Chow, S. T., Lim, B. S., Tickner, J., Zhao, J., Chung, R., Su, Y.-W., Zhang, G., Erber, W., Xian, C. J., Rosen, V., and Xu, J. (2014). EGFL7 is expressed in bone microenvironment and promotes angiogenesis via ERK, STAT3, and integrin signaling cascades. *Journal of Cellular Physiology*, 230(1):82–94.

[Choi et al., 1998] Choi, K., Kennedy, M., Kazarov, A., Papadimitriou, J., and Keller, G. (1998). A common precursor for hematopoietic and endothelial cells. *Development*, 125(4):725–732.

[Chong et al., 2014] Chong, J. J., Forte, E., and Harvey, R. P. (2014). Developmental origins and lineage descendants of endogenous adult cardiac progenitor cells. *Stem Cell Research*, 13(3, Part B):592 – 614.

[Costa et al., 2012] Costa, G., Mazan, A., Gandillet, A., Pearson, S., Lacaud, G., and Kouskoff, V. (2012). SOX7 regulates the expression of VE-cadherin in the haemogenic endothelium at the onset of haematopoietic development. *Development*, 139(9):1587–1598.

[Cowan et al., 2014] Cowan, J., Tariq, M., and Ware, S. M. (2014). Genetic and functional analyses of ZIC3 variants in congenital heart disease. *Human Mutation*, 35(1):66–75.

[Cowley and Oakey, 2012] Cowley, M. and Oakey, R. J. (2012). Resetting for the next generation. *Molecular Cell*, 48(6):819 – 821.

[Creyghton et al., 2010] Creyghton, M. P., Cheng, A. W., Welstead, G. G., Kooistra, T., Carey, B. W., Steine, E. J., Hanna, J., Lodato, M. A., Frampton, G. M., Sharp, P. A., Boyer, L. A., Young, R. A., and Jaenisch, R. (2010). Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proceedings of the National Academy of Sciences*, 107(50):21931–21936.

[Davies et al., 2005] Davies, W., Isles, A. R., and Wilkinson, L. S. (2005). Imprinted gene expression in the brain. *Neuroscience and Biobehavioral Reviews*, 29(3):421 – 430.

[de Felipe and Ryan, 2004] de Felipe, P. and Ryan, M. D. (2004). Targeting of proteins derived from self-processing polyproteins containing multiple signal sequences. *Traffic*, 5(8):616–626.

[de la Pompa et al., 1998] de la Pompa, J. L., Timmerman, L. A., Takimoto, H., Yoshida, H., Elia, A. J., Samper, E., Potter, J., Wakeham, A., Marengere, L., Langille, B. L., Crabtree, G. R., and Mak, T. W. (1998). Role of the NF-ATc transcription factor in morphogenesis of cardiac valves and septum. *Nature*, 392(6672):182–186.

[de Vlaming et al., 2012] de Vlaming, A., Sauls, K., Hajdu, Z., Visconti, R. P., Mehesz, A. N., Levine, R. A., Slaugenhaupt, S. A., Hagège, A., Chester, A. H., Markwald, R. R., and Norris, R. A. (2012). Atrioventricular valve development: New perspectives on an old theme. *Differentiation*, 84(1):103 – 116.

[Deaton and Bird, 2011] Deaton, A. M. and Bird, A. (2011). CpG islands and the regulation of transcription. *Genes and Development*, 25(10):1010–1022.

[Deaton et al., 2011] Deaton, A. M., Webb, S., Kerr, A. R., Illingworth, R. S., Guy, J., Andrews, R., and Bird, A. (2011). Cell type–specific DNA methylation at intragenic CpG islands in the immune system. *Genome Research*, 21(7):1074–1086.

[Dejana and Orsenigo, 2013] Dejana, E. and Orsenigo, F. (2013). Endothelial adherens junctions at a glance. *Journal of Cell Science*, 126(12):2545–2549.

[DeLaughter et al., 2011] DeLaughter, D. M., Saint-Jean, L., Baldwin, H. S., and Barnett, J. V. (2011). What chick and mouse models have taught us about the role of the endocardium in congenital heart disease. *Birth Defects Research Part A: Clinical and Molecular Teratology*, 91(6):511–525.

[Deleuze et al., 2007] Deleuze, V., Chalhoub, E., El-Hajj, R., Dohet, C., Le Clech, M., Couraud, P.-O., Huber, P., and Mathieu, D. (2007). TAL-1/SCL and its partners E47 and LMO2 up-regulate VE-cadherin expression in endothelial cells. *Molecular and Cellular Biology*, 27(7):2687–2697.

[Denis et al., 2011] Denis, H., Ndlovu, M. N., and Fuks, F. (2011). Regulation of mammalian DNA methyltransferases: a route to new mechanisms. *EMBO Reports*, 12(7):647–656.

[Doll and Burlingame, 2015] Doll, S. and Burlingame, A. L. (2015). Mass spectrometry-based detection and assignment of protein posttranslational modifications. *ACS Chemical Biology*, 10(1):63–71. PMID: 25541750.

[Domínguez et al., 2012] Domínguez, J. N., Meilhac, S. M., Bland, Y. S., Buckingham, M. E., and Brown, N. A. (2012). Asymmetric fate of the posterior part of the second heart field results in unexpected left/right contributions to both poles of the heart. *Circulation Research*, 111(10):1323–1335.

[Dominguez et al., 2007] Dominguez, M. G., Hughes, V. C., Pan, L., Simmons, M., Daly, C., Anderson, K., Noguera-Troise, I., Murphy, A. J., Valenzuela, D. M., Davis, S., Thurston, G., Yancopoulos, G. D., and Gale, N. W. (2007). Vascular endothelial tyrosine phosphatase (VE-PTP)-null mice undergo vasculogenesis but die embryonically because of defects in angiogenesis. *Proceedings of the National Academy of Sciences*, 104(9):3243–3248.

[Doppler et al., 2014] Doppler, S. A., Werner, A., Barz, M., Lahm, H., Deutsch, M.-A., Dreßen, M., Schiemann, M., Voss, B., Gregoire, S., Kuppusamy, R., Wu, S. M., Lange, R., and Krane, M. (2014). Myeloid zinc finger 1 (Mzf1) differentially modulates murine cardiogenesis by interacting with an Nkx2.5 cardiac enhancer. *PLoS ONE*, 9(12):e113775.

[Dumont et al., 1994] Dumont, D. J., Gradwohl, G., Fong, G. H., Puri, M. C., Gertsenstein, M., Auerbach, A., and Breitman, M. L. (1994). Dominant-negative and targeted null mutations in the endothelial receptor tyrosine kinase, tek, reveal a critical role in vasculogenesis of the embryo. *Genes & Development*, 8(16):1897–1909.

[Dyer and Kirby, 2009] Dyer, L. A. and Kirby, M. L. (2009). The role of secondary heart field in cardiac development. *Developmental Biology*, 336(2):137 – 144.

[Edelstein et al., 2005] Edelstein, L. C., Pan, A., and Collins, T. (2005). Chromatin modification and the endothelial-specific activation of the e-selectin gene. *Journal of Biological Chemistry*, 280(12):11192–11202.

[Farkas et al., 2013] Farkas, C., Martins, C. P., Escobar, D., Hepp, M. I., Donner, D. B., Castro, A. F., Evan, G., Gutiérrez, J. L., Warren, R., and Pincheira, R. (2013). Wild type p53 transcriptionally represses the SALL2 transcription factor under genotoxic stress. *PLoS ONE*, 8(9):e73817.

[Ferdous et al., 2009] Ferdous, A., Caprioli, A., Iacovino, M., Martin, C. M., Morris, J., Richardson, J. A., Latif, S., Hammer, R. E., Harvey, R. P., Olson, E. N., Kyba, M., and Garry, D. J. (2009). Nkx2–5 transactivates the Ets-related protein 71 gene and specifies an endothelial/endocardial fate in the developing embryo. *Proceedings of the National Academy of Sciences*, 106(3):814–819.

[Filippova et al., 1996] Filippova, G. N., Fagerlie, S., Klenova, E. M., Myers, C., Dehner, Y., Goodwin, G., Neiman, P. E., Collins, S. J., and Lobanenkov, V. V. (1996). An exceptionally conserved transcriptional repressor, CTCF, employs different combinations of zinc fingers to bind diverged promoter sequences of avian and mammalian c-myc oncogenes. *Molecular and Cellular Biology*, 16(6):2802–13.

[Fish et al., 2005] Fish, J. E., Matouk, C. C., Rachlis, A., Lin, S., Tai, S. C., D'Abreo, C., and Marsden, P. A. (2005). The expression of endothelial nitric-oxide synthase is controlled by a cell-specific histone code. *Journal of Biological Chemistry*, 280(26):24824–24838.

[Francois et al., 2010] Francois, M., Koopman, P., and Beltrame, M. (2010). SoxF genes: Key players in the development of the cardio-vascular system. *The International Journal of Biochemistry & Cell Biology*, 42(3):445 – 448.

[Fu et al., 2009] Fu, X., Fu, N., Guo, S., Yan, Z., Xu, Y., Hu, H., Menzel, C., Chen, W., Li, Y., Zeng, R., and Khaitovich, P. (2009). Estimating accuracy of RNA-Seq and microarrays with proteomics. *BMC Genomics*, 10(1):161.

[Gardiner-Garden and Frommer, 1987] Gardiner-Garden, M. and Frommer, M. (1987). CpG islands in vertebrate genomes. *Journal of Molecular Biology*, 196(2):261 – 282.

[Giardine et al., 2005] Giardine, B., Riemer, C., Hardison, R. C., Burhans, R., Elnitski, L., Shah, P., Zhang, Y., Blankenberg, D., Albert, I., Taylor, J., Miller, W., Kent, W. J., and Nekrutenko, A. (2005). Galaxy: A platform for interactive large-scale genome analysis. *Genome Research*, 15(10):1451–1455.

[Gifford et al., 2013] Gifford, C. A., Ziller, M. J., Gu, H., Trapnell, C., Donaghey, J., Tsankov, A., Shalek, A. K., Kelley, D. R., Shishkin, A. A., Issner, R., Zhang, X., Coyne, M., Fostel, J. L., Holmes, L., Meldrim, J., Guttman, M., Epstein, C., Park, H., Kohlbacher, O., Rinn, J., Gnirke, A., Lander, E. S., Bernstein, B. E., and Meissner,

A. (2013). Transcriptional and epigenetic dynamics during specification of human embryonic stem cells. *Cell*, 153(5):1149 – 1163.

[Gilbert, 2003] Gilbert, S. (2003). *Developmental Biology*. Sinauer Associates.

[Goecks et al., 2010] Goecks, J., Nekrutenko, A., Taylor, J., and Team, T. G. (2010). Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biology*, 11(8):R86.

[Goff et al., 2012] Goff, L., Trapnell, C., and Kelley, D. (2012). *cummeRbund: Analysis, exploration, manipulation, and visualization of Cufflinks high-throughput sequencing data*. R package version 2.6.1.

[Goldberg et al., 2007] Goldberg, A. D., Allis, C. D., and Bernstein, E. (2007). Epigenetics: A landscape takes shape. *Cell*, 128(4):635 – 638.

[Goll et al., 2006] Goll, M. G., Kirpekar, F., Maggert, K. A., Yoder, J. A., Hsieh, C.-L., Zhang, X., Golic, K. G., Jacobsen, S. E., and Bestor, T. H. (2006). Methylation of tRNAAsp by the DNA methyltransferase homolog Dnmt2. *Science*, 311(5759):395–398.

[Guelen et al., 2008] Guelen, L., Pagie, L., Brasset, E., Meuleman, W., Faza, M. B., Talhout, W., Eussen, B. H., de Klein, A., Wessels, L., de Laat, W., and van Steensel, B. (2008). Domain organization of human chromosomes revealed by mapping of nuclear lamina interactions. *Nature*, 453(7197):948–951.

[Gullerova and Proudfoot, 2008] Gullerova, M. and Proudfoot, N. J. (2008). Cohesin complex promotes transcriptional termination between convergent genes in S. pombe. *Cell*, 132(6):983–995.

[Guo et al., 2011a] Guo, C., Yoon, H. S., Franklin, A., Jain, S., Ebert, A., Cheng, H.-L., Hansen, E., Despo, O., Bossen, C., Vettermann, C., Bates, J. G., Richards, N., Myers, D., Patel, H., Gallagher, M., Schlissel, M. S., Murre, C., Busslinger, M., Giallourakis, C. C., and Alt, F. W. (2011a). CTCF-binding elements mediate control of V(D)J recombination. *Nature*, 477(7365):424–430.

[Guo et al., 2014] Guo, F., Li, X., Liang, D., Li, T., Zhu, P., Guo, H., Wu, X., Wen, L., Gu, T.-P., Hu, B., Walsh, C. P., Li, J., Tang, F., and Xu, G.-L. (2014). Active and passive demethylation of male and female pronuclear DNA in the mammalian zygote. *Cell Stem Cell*, 15(4):447 – 458.

[Guo et al., 2011b] Guo, J. U., Ma, D. K., Mo, H., Ball, M. P., Jang, M.-H., Bonaguidi, M. A., Balazer, J. A., Eaves, H. L., Xie, B., Ford, E., Zhang, K., Ming, G.-l., Gao, Y., and Song, H. (2011b). Neuronal activity modifies the dna methylation landscape in the adult brain. *Nat Neurosci*, 14(10):1345–1351.

[Guo et al., 2011c] Guo, J. U., Su, Y., Zhong, C., li Ming, G., and Song, H. (2011c). Hydroxylation of 5-methylcytosine by TET1 promotes active DNA demethylation in the adult brain. *Cell*, 145(3):423 – 434.

[Hadjur et al., 2009] Hadjur, S., Williams, L. M., Ryan, N. K., Cobb, B. S., Sexton, T., Fraser, P., Fisher, A. G., and Merkenschlager, M. (2009). Cohesins form chromosomal cis-interactions at the developmentally regulated IFNG locus. *Nature*, 460(7253):410–413.

[Hajkova, 2010] Hajkova, P. (2010). Epigenetic reprogramming — taking a lesson from the embryo. *Current Opinion in Cell Biology*, 22(3):342 – 350.

[Hajkova et al., 2008] Hajkova, P., Ancelin, K., Waldmann, T., Lacoste, N., Lange, U. C., Cesari, F., Lee, C., Almouzni, G., Schneider, R., and Surani, M. A. (2008). Chromatin dynamics during epigenetic reprogramming in the mouse germ line. *Nature*, 452(7189):877–881.

[Halg and Graham, 1991] Halg, D. and Graham, C. (1991). Genomic imprinting and the strange case of the insulin-like growth factor II receptor. *Cell*, 64(6):1045 – 1046.

[Handoko et al., 2011] Handoko, L., Xu, H., Li, G., Ngan, C. Y., Chew, E., Schnapp, M., Lee, C. W. H., Ye, C., Ping, J. L. H., Mulawadi, F., Wong, E., Sheng, J., Zhang, Y., Poh, T., Chan, C. S., Kunarso, G., Shahab, A., Bourque, G., Cacheux-Rataboul, V., Sung, W.-K., Ruan, Y., and Wei, C.-L. (2011). Ctcf-mediated functional chromatin interactome in pluripotent cells. *Nat Genet*, 43(7):630–638.

[Hansen et al., 2012] Hansen, K., Langmead, B., and Irizarry, R. (2012). BSmooth: from whole genome bisulfite sequencing reads to differentially methylated regions. *Genome Biology*, 13(10):R83.

[Hansen et al., 2011] Hansen, K. D., Timp, W., Bravo, H. C., Sabunciyan, S., Langmead, B., McDonald, O. G., Wen, B., Wu, H., Liu, Y., Diep, D., Briem, E., Zhang, K., Irizarry,

R. A., and Feinberg, A. P. (2011). Increased methylation variation in epigenetic domains across cancer types. *Nat Genet*, 43(8):768–775.

[Hansen et al., 2008] Hansen, K. H., Bracken, A. P., Pasini, D., Dietrich, N., Gehani, S. S., Monrad, A., Rappsilber, J., Lerdrup, M., and Helin, K. (2008). A model for transmission of the H3K27me3 epigenetic mark. *Nature Cell Biology*, 10(11):1291–1300.

[Harmon and Nakano, 2013] Harmon, A. W. and Nakano, A. (2013). Nkx2-5 lineage tracing visualizes the distribution of second heart field-derived aortic smooth muscle. *genesis*, 51(12):862–869.

[Harris and Black, 2010] Harris, I. and Black, B. (2010). Development of the endocardium. *Pediatric Cardiology*, 31(3):391–399.

[Harris et al., 2010] Harris, T. A., Yamakuchi, M., Kondo, M., Oettgen, P., and Lowenstein, C. J. (2010). Ets-1 and Ets-2 regulate the expression of MicroRNA-126 in endothelial cells. *Arteriosclerosis, Thrombosis, and Vascular Biology*, 30(10):1990–1997.

[Hikichi et al., 2003] Hikichi, T., Kohda, T., KanekoIshino, T., and Ishino, F. (2003). Imprinting regulation of the murine Meg1/Grb10 and human GRB10 genes; roles of brain–specific promoters and mouse–specific CTCF–binding sites. *Nucleic Acids Research*, 31(5):1398–1406.

[Hinton and Yutzey, 2011] Hinton, R. B. and Yutzey, K. E. (2011). Heart valve structure and function in development and disease. *Annual Review of Physiology*, 73(1):29–46.

[Hoffman and Kaplan, 2002] Hoffman, J. I. and Kaplan, S. (2002). The incidence of congenital heart disease. *Journal of the American College of Cardiology*, 39(12):1890 – 1900.

[Hoffman et al., 2004] Hoffman, J. I., Kaplan, S., and Liberthson, R. R. (2004). Prevalence of congenital heart disease. *American Heart Journal*, 147(3):425 – 439.

[Huang et al., 2009] Huang, J., Min Lu, M., Cheng, L., Yuan, L.-J., Zhu, X., Stout, A. L., Chen, M., Li, J., and Parmacek, M. S. (2009). Myocardin is required for cardiomyocyte survival and maintenance of heart function. *Proceedings of the National Academy of Sciences*, 106(44):18734–18739.

[Huang et al., 2005] Huang, X., Brown, C., Ni, W., Maynard, E., Rigby, A. C., and Oettgen, P. (2005). Critical role for the Ets transcription factor ELF-1 in the development of tumor angiogenesis. *Blood*, 107(8):3153–3160.

[Illingworth et al., 2010] Illingworth, R. S., Gruenewald-Schneider, U., Webb, S., Kerr, A. R. W., James, K. D., Turner, D. J., Smith, C., Harrison, D. J., Andrews, R., and Bird, A. P. (2010). Orphan CpG islands identify numerous conserved promoters in the mammalian genome. *PLoS Genetics*, 6(9):e1001134.

[Illumina, 2011] Illumina (2011). Quality scores for next-generation sequencing. *online resource*. http://res.illumina.com/documents/products/technotes/technote-q-scores.pdf.

[Isensee et al., 2008] Isensee, J., Witt, H., Pregla, R., Hetzer, R., Regitz-Zagrosek, V., and Ruiz Noppinger, P. (2008). Sexually dimorphic gene expression in the heart of mice and men. *Journal of Molecular Medicine*, 86(1):61–74.

[Islam et al., 2014] Islam, S., Zeisel, A., Joost, S., La Manno, G., Zajac, P., Kasper, M., Lonnerberg, P., and Linnarsson, S. (2014). Quantitative single-cell RNA-seq with unique molecular identifiers. *Nature Methods*, 11(2):163–166.

[Ito et al., 2011] Ito, S., Shen, L., Dai, Q., Wu, S. C., Collins, L. B., Swenberg, J. A., He, C., and Zhang, Y. (2011). Tet proteins can convert 5-methylcytosine to 5-formylcytosine and 5-carboxylcytosine. *Science*, 333(6047):1300–1303.

[Jakobsson et al., 2010] Jakobsson, L., Franco, C. A., Bentley, K., Collins, R. T., Ponsioen, B., Aspalter, I. M., Rosewell, I., Busse, M., Thurston, G., Medvinsky, A., Schulte-Merker, S., and Gerhardt, H. (2010). Endothelial cells dynamically compete for the tip cell position during angiogenic sprouting. *Nature Cell Biology*, 12(10):943–953.

[Jones, 2012] Jones, P. A. (2012). Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nature Review Genetics*, 13(7):484–492.

[Kagey et al., 2010] Kagey, M. H., Newman, J. J., Bilodeau, S., Zhan, Y., Orlando, D. A., van Berkum, N. L., Ebmeier, C. C., Goossens, J., Rahl, P. B., Levine, S. S., Taatjes, D. J., Dekker, J., and Young, R. A. (2010). Mediator and cohesin connect gene expression and chromatin architecture. *Nature*, 467(7314):430–435.

[Kanamori et al., 2004] Kanamori, M., Konno, H., Osato, N., Kawai, J., Hayashizaki, Y., and Suzuki, H. (2004). A genome-wide and nonredundant mouse transcription factor database. *Biochemical and Biophysical Research Communications*, 322(3):787 – 793.

[Kanduri et al., 2000] Kanduri, C., Pant, V., Loukinov, D., Pugacheva, E., Qi, C.-F., Wolffe, A., Ohlsson, R., and Lobanenkov, V. V. (2000). Functional association of CTCF with the insulator upstream of the H19 gene is parent of origin-specific and methylation-sensitive. *Current Biology*, 10(14):853 – 856.

[Kaneda et al., 2009] Kaneda, R., Takada, S., Yamashita, Y., Choi, Y. L., Nonaka-Sarukawa, M., Soda, M., Misawa, Y., Isomura, T., Shimada, K., and Mano, H. (2009). Genome-wide histone methylation profile for heart failure. *Genes to Cells*, 14(1):69–77.

[Kattman et al., 2006] Kattman, S. J., Huber, T. L., and Keller, G. M. (2006). Multipotent Flk-1+ cardiovascular progenitor cells give rise to the cardiomyocyte, endothelial, and vascular smooth muscle lineages. *Developmental Cell*, 11(5):723–732.

[Keverne, 1997] Keverne, E. B. (1997). Genomic imprinting in the brain. *Current Opinion in Neurobiology*, 7(4):463 – 468.

[Keverne et al., 1996] Keverne, E. B., Fundele, R., Narasimha, M., Barton, S. C., and Surani, M. (1996). Genomic imprinting and the differential roles of parental genomes in brain development. *Developmental Brain Research*, 92(1):91 – 100.

[Kim et al., 2013] Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R., and Salzberg, S. (2013). Tophat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biology*, 14(4):R36.

[Kim and Kim, 2014] Kim, J. H. and Kim, N. (2014). Regulation of NFATc1 in osteoclast differentiation. *J Bone Metab*, 21(4):233–241.

[Kim et al., 2007] Kim, T. H., Abdullaev, Z. K., Smith, A. D., Ching, K. A., Loukinov, D. I., Green, R. D., Zhang, M. Q., Lobanenkov, V. V., and Ren, B. (2007). Analysis of the vertebrate insulator protein CTCF-binding sites in the human genome. *Cell*, 128(6):1231 – 1245.

[Kohli and Zhang, 2013] Kohli, R. M. and Zhang, Y. (2013). TET enzymes, TDG and the dynamics of DNA demethylation. *Nature*, 502(7472):472–479.

[Krause et al., 2013] Krause, B. J., Costello, P. M., Muñoz-Urrutia, E., Lillycrop, K. A., Hanson, M. A., and Casanello, P. (2013). Role of DNA methyltransferase 1 on the altered eNOS expression in human umbilical endothelium from intrauterine growth restricted fetuses. *Epigenetics*, 8(9):944–952.

[Krueger and Andrews, 2011] Krueger, F. and Andrews, S. R. (2011). Bismark: a flexible aligner and methylation caller for bisulfite-seq applications. *Bioinformatics*, 27(11):1571–1572.

[Kuzmin et al., 2005] Kuzmin, I., Geil, L., Gibson, L., Cavinato, T., Loukinov, D., Lobanenkov, V., and Lerman, M. I. (2005). Transcriptional regulator CTCF controls human interleukin 1 receptor-associated kinase 2 promoter. *Journal of Molecular Biology*, 346(2):411 – 422.

[LaHaye et al., 2014] LaHaye, S., Lincoln, J., and Garg, V. (2014). Genetics of valvular heart disease. *Current Cardiology Reports*, 16(6):487–495.

[Landry et al., 2008] Landry, J.-R., Kinston, S., Knezevic, K., de Bruijn, M. F., Wilson, N., Nottingham, W. T., Peitz, M., Edenhofer, F., Pimanda, J. E., Ottersbach, K., and Göttgens, B. (2008). Runx genes are direct targets of Scl/Tal1 in the yolk sac and fetal liver. *Blood*, 111(6):3005–3014.

[Langmead and Salzberg, 2012] Langmead, B. and Salzberg, S. L. (2012). Fast gapped-read alignment with bowtie 2. *Nat Meth*, 9(4):357–359.

[Langmead et al., 2009] Langmead, B., Trapnell, C., Pop, M., and Salzberg, S. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology*, 10(3):R25.

[Larson et al., 2014] Larson, A. M., Shinnick, J. E., Shaaya, E. A., Thiele, E. A., and Thibert, R. L. (2014). Angelman syndrome in adulthood. *American Journal of Medical Genetics Part A*, 167A(2):331–344.

[Leirgul et al., 2014] Leirgul, E., Fomina, T., Brodwall, K., Greve, G., Holmstrøm, H., Vollset, S. E., Tell, G. S., and Øyen, N. (2014). Birth prevalence of congenital heart defects in norway 1994-2009—a nationwide study. *American Heart Journal*, 168(6):956 – 964.

[Lepore et al., 2006] Lepore, J. J., Mericko, P. A., Cheng, L., Lu, M. M., Morrisey, E. E., and Parmacek, M. S. (2006). Gata-6 regulates semaphorin 3c and is required in cardiac neural crest for cardiovascular morphogenesis. *The Journal of Clinical Investigation*, 116(4):929–939.

[Lewis et al., 2014] Lewis, M. W., Brant, J. O., Kramer, J. M., Moss, J. I., Yang, T. P., Hansen, P., Williams, R. S., and Resnick, J. L. (2014). Angelman syndrome imprinting center encodes a transcriptional promoter. *Proceedings of the National Academy of Sciences*.

[Li et al., 2009] Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and Subgroup, . G. P. D. P. (2009). The sequence alignment/map format and samtools. *Bioinformatics*, 25(16):2078–2079.

[Lin et al., 2011] Lin, S., Ferguson-Smith, A. C., Schultz, R. M., and Bartolomei, M. S. (2011). Nonallelic transcriptional roles of CTCF and cohesins at imprinted loci. *Molecular and Cellular Biology*, 31(15):3094–3104.

[Linask, 1992] Linask, K. K. (1992). N-cadherin localization in early heart development and polar expression of Na^+, K^+-ATPase, and integrin during pericardial coelom formation and epithelialization of the differentiating myocardium. *Developmental Biology*, 151(1):213–224.

[Lincoln and Garg, 2014] Lincoln, J. and Garg, V. (2014). Etiology of valvular heart disease. *Circulation Journal*, 78(8):1801–1807.

[Lister et al., 2009] Lister, R., Pelizzola, M., Dowen, R. H., Hawkins, R. D., Hon, G., Tonti-Filippini, J., Nery, J. R., Lee, L., Ye, Z., Ngo, Q.-M., Edsall, L., Antosiewicz-Bourget, J., Stewart, R., Ruotti, V., Millar, A. H., Thomson, J. A., Ren, B., and Ecker, J. R. (2009). Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature*, 462(7271):315–322.

[Liu and Feng, 2012] Liu, Y. and Feng, Q. (2012). NOing the heart: Role of nitric oxide synthase-3 in heart development. *Differentiation*, 84(1):54–61.

[Lobanenkov et al., 1990] Lobanenkov, V., Nicolas, R., Adler, V., Paterson, H., Klenova, E., Polotskaja, A., and Goodwin, G. (1990). A novel sequence-specific DNA binding

protein which interacts with three regularly spaced direct repeats of the CCCTC-motif in the 5'-flanking sequence of the chicken c-myc gene. *Oncogene*, 5(12):1743–1753.

[Luger et al., 1997] Luger, K., Mader, A. W., Richmond, R. K., Sargent, D. F., and Richmond, T. J. (1997). Crystal structure of the nucleosome core particle at 2.8A resolution. *Nature*, 389(6648):251–260.

[Lyck et al., 2007] Lyck, L., Krøigård, T., and Finsen, B. (2007). Unbiased cell quantification reveals a continued increase in the number of neocortical neurones during early post-natal development in mice. *European Journal of Neuroscience*, 26(7):1749–1764.

[Martinowich et al., 2003] Martinowich, K., Hattori, D., Wu, H., Fouse, S., He, F., Hu, Y., Fan, G., and Sun, Y. E. (2003). DNA methylation-related chromatin remodeling in activity-dependent Bdnf gene regulation. *Science*, 302(5646):890–893.

[Mathiyalagan et al., 2010] Mathiyalagan, P., Chang, L., Du, X.-J., and El-Osta, A. (2010). Cardiac ventricular chambers are epigenetically distinguishable. *Cell Cycle*, 9(3):612–617.

[Maunakea et al., 2013] Maunakea, A. K., Chepelev, I., Cui, K., and Zhao, K. (2013). Intragenic DNA methylation modulates alternative splicing by recruiting MeCP2 to promote exon recognition. *Cell Research*, 23(11):1256–1269.

[McBride et al., 2008] McBride, K. L., Riley, M. F., Zender, G. A., Fitzgerald-Butt, S. M., Towbin, J. A., Belmont, J. W., and Cole, S. E. (2008). Notch1 mutations in individuals with left ventricular outflow tract malformations reduce ligand-induced signaling. *Human Molecular Genetics*, 17(18):2886–2893.

[McGrath and Solter, 1984] McGrath, J. and Solter, D. (1984). Completion of mouse embryogenesis requires both the maternal and paternal genomes. *Cell*, 37(1):179 – 183.

[Meissner et al., 2005] Meissner, A., Gnirke, A., Bell, G. W., Ramsahoye, B., Lander, E. S., and Jaenisch, R. (2005). Reduced representation bisulfite sequencing for comparative high-resolution DNA methylation analysis. *Nucleic Acids Research*, 33(18):5868–5877.

[Mikawa and Hurtado, 2007] Mikawa, T. and Hurtado, R. (2007). Development of the cardiac conduction system. *Seminars in Cell and Developmental Biology*, 18(1):90 – 100. Model Systems for the Study of Cardiovascular Development and Disease.

[Milgrom-Hoffman et al., 2011] Milgrom-Hoffman, M., Harrelson, Z., Ferrara, N., Zelzer, E., Evans, S. M., and Tzahor, E. (2011). The heart endocardium is derived from vascular endothelial progenitors. *Development*, 138(21):4777–4787.

[Misfeldt et al., 2009] Misfeldt, A. M., Boyle, S. C., Tompkins, K. L., Bautch, V. L., Labosky, P. A., and Baldwin, H. S. (2009). Endocardial cells are a distinct endothelial lineage derived from Flk1+ multipotent cardiovascular progenitors. *Developmental Biology*, 333(1):78 – 89.

[Mjaatvedt et al., 1987] Mjaatvedt, C., Lepera, R., and Markwald, R. (1987). Myocardial specificity for initiating endothelial-mesenchymal cell transition in embryonic chick heart correlates with a particulate distribution of fibronectin. *Developmental Biology*, 119(1):59 – 67.

[Moore and Haig, 1991] Moore, T. and Haig, D. (1991). Genomic imprinting in mammalian development: a parental tug-of-war. *Trends in Genetics*, 7(2):45 – 49.

[Moorman et al., 2003] Moorman, A., Webb, S., Brown, N. A., Lamers, W., and Anderson, R. H. (2003). Development of the heart: (1) formation of the cardiac chambers and arterial trunks. *Heart*, 89(7):806–814.

[Moretti et al., 2006] Moretti, A., Caron, L., Nakano, A., Lam, J. T., Bernshausen, A., Chen, Y., Qyang, Y., Bu, L., Sasaki, M., Martin-Puig, S., Sun, Y., Evans, S. M., Laugwitz, K.-L., and Chien, K. R. (2006). Multipotent embryonic Isl1+ progenitor cells lead to cardiac, smooth muscle, and endothelial cell diversification. *Cell*, 127(6):1151–1165.

[Morison et al., 2005] Morison, I. M., Ramsay, J. P., and Spencer, H. G. (2005). A census of mammalian imprinting. *Trends in Genetics*, 21(8):457 – 465.

[Mortazavi et al., 2008] Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L., and Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods*, 5(7):621–628.

[Murphy et al., 2001] Murphy, S. K., Wylie, A. A., and Jirtle, R. L. (2001). Imprinting of peg3, the human homologue of a mouse gene involved in nurturing behavior. *Genomics*, 71(1):110 – 117.

[Nadeau et al., 2010] Nadeau, M., Georges, R. O., Laforest, B., Yamak, A., Lefebvre, C., Beauregard, J., Paradis, P., Bruneau, B. G., Andelfinger, G., and Nemer, M. (2010). An endocardial pathway involving Tbx5, Gata4, and Nos3 required for atrial septum formation. *Proceedings of the National Academy of Sciences*, 107(45):19356–19361.

[Narumiya et al., 2007] Narumiya, H., Hidaka, K., Shirai, M., Terami, H., Aburatani, H., and Morisaki, T. (2007). Endocardiogenesis in embryoid bodies: Novel markers identified by gene expression profiling. *Biochemical and Biophysical Research Communications*, 357(4):896 – 902.

[Nativio et al., 2009] Nativio, R., Wendt, K. S., Ito, Y., Huddleston, J. E., Uribe-Lewis, S., Woodfine, K., Krueger, C., Reik, W., Peters, J.-M., and Murrell, A. (2009). Cohesin is required for higher-order chromatin conformation at the imprinted IGF2-H19 locus. *PLoS Genet*, 5(11):e1000739.

[Neph et al., 2012] Neph, S., Kuehn, M. S., Reynolds, A. P., Haugen, E., Thurman, R. E., Johnson, A. K., Rynes, E., Maurano, M. T., Vierstra, J., Thomas, S., Sandstrom, R., Humbert, R., and Stamatoyannopoulos, J. A. (2012). Bedops: high-performance genomic feature operations. *Bioinformatics*, 28(14):1919–1920.

[Nichol and Stuhlmann, 2011] Nichol, D. and Stuhlmann, H. (2011). Egfl7: a unique angiogenic signaling factor in vascular development and disease. *Blood*, 119(6):1345–1352.

[Nie and Bronner, 2015] Nie, S. and Bronner, M. E. (2015). Dual developmental role of transcriptional regulator ets1 in xenopus cardiac neural crest vs. heart mesoderm. *Cardiovascular Research*, 106(1):67–75.

[Nishikawa et al., 1998] Nishikawa, S., Nishikawa, S., Hirashima, M., Matsuyoshi, N., and Kodama, H. (1998). Progressive lineage analysis by cell sorting and culture identifies FLK1+ VE-cadherin+ cells at a diverging point of endothelial and hemopoietic lineages. *Development*, 125(9):1747–1757.

[Niwa et al., 2009] Niwa, H., Ogawa, K., Shimosato, D., and Adachi, K. (2009). A parallel circuit of LIF signalling pathways maintains pluripotency of mouse ES cells. *Nature*, 460(7251):118–122.

[Nix et al., 2008] Nix, D., Courdy, S., and Boucher, K. (2008). Empirical methods for controlling false positives and estimating confidence in ChIP-Seq peaks. *BMC Bioinformatics*, 9(1):523.

[Ocampo-Hafalla and Uhlmann, 2011] Ocampo-Hafalla, M. T. and Uhlmann, F. (2011). Cohesin loading and sliding. *Journal of Cell Science*, 124(5):685–691.

[Okada et al., 2014] Okada, Y., Funahashi, N., Tanaka, T., Nishiyama, Y., Yuan, L., Shirakura, K., Turjman, A. S., Kano, Y., Naruse, H., Suzuki, A., Sakai, M., Zhixia, J., Kitajima, K., Ishimoto, K., Hino, N., Kondoh, M., Mukai, Y., Nakagawa, S., García-Cardeña, G., Aird, W. C., and Doi, T. (2014). Endothelial cell–specific expression of roundabout 4 is regulated by differential dna methylation of the proximal promoter. *Arteriosclerosis, Thrombosis, and Vascular Biology*, 34(7):1531–1538.

[Onn et al., 2008] Onn, I., Heidinger-Pauli, J. M., Guacci, V., Ünal, E., and Koshland, D. E. (2008). Sister chromatid cohesion: A simple concept with a complex reality. *Annual Review of Cell and Developmental Biology*, 24(1):105–129.

[Paige et al., 2012] Paige, S. L., Thomas, S., Stoick-Cooper, C. L., Wang, H., Maves, L., Sandstrom, R., Pabon, L., Reinecke, H., Pratt, G., Keller, G., Moon, R. T., Stamatoyannopoulos, J., and Murry, C. E. (2012). A temporal chromatin signature in human embryonic stem cells identifies regulators of cardiac development. *Cell*, 151(1):221 – 232.

[Parelho et al., 2008] Parelho, V., Hadjur, S., Spivakov, M., Leleu, M., Sauer, S., Gregson, H. C., Jarmuz, A., Canzonetta, C., Webster, Z., Nesterova, T., Cobb, B. S., Yokomori, K., Dillon, N., Aragon, L., Fisher, A. G., and Merkenschlager, M. (2008). Cohesins functionally associate with CTCF on mammalian chromosome arms. *Cell*, 132(3):422 – 433.

[Patten et al., 2014] Patten, M. M., Ross, L., Curley, J. P., Queller, D. C., Bonduriansky, R., and Wolf, J. B. (2014). The evolution of genomic imprinting: theories, predictions and empirical tests. *Heredity*, 113(2):119–128.

[Pearson et al., 1997] Pearson, W. R., Wood, T., Zhang, Z., and Miller, W. (1997). Comparison of DNA sequences with protein sequences. *Genomics*, 46(1):24 – 36.

[Peng and Jahroudi, 2003] Peng, Y. and Jahroudi, N. (2003). The nfy transcription factor inhibits von willebrand factor promoter activation in non-endothelial cells through recruitment of histone deacetylases. *Journal of Biological Chemistry*, 278(10):8385–8394.

[Pepke et al., 2009] Pepke, S., Wold, B., and Mortazavi, A. (2009). Computation for ChIP-seq and RNA-seq studies. *Nat Meth*, 6(11s):S22–S32.

[Phillips and Corces, 2009] Phillips, J. E. and Corces, V. G. (2009). CTCF: Master weaver of the genome. *Cell*, 137(7):1194–1211.

[Prickett et al., 2013] Prickett, A. R., Barkas, N., McCole, R. B., Hughes, S., Amante, S. M., Schulz, R., and Oakey, R. J. (2013). Genome-wide and parental allele-specific analysis of CTCF and cohesin dna binding in mouse brain reveals a tissue-specific binding pattern and an association with imprinted differentially methylated regions. *Genome Research*, 23(10):1624–1635.

[Puri et al., 1999] Puri, M., Partanen, J., Rossant, J., and Bernstein, A. (1999). Interaction of the TEK and TIE receptor tyrosine kinases during cardiovascular development. *Development*, 126(20):4569–4580.

[Quail et al., 2008] Quail, M. A., Kozarewa, I., Smith, F., Scally, A., Stephens, P. J., Durbin, R., Swerdlow, H., and Turner, D. J. (2008). A large genome center's improvements to the Illumina sequencing system. *Nature Methods*, 5(12):1005–1010.

[Quinlan and Hall, 2010] Quinlan, A. R. and Hall, I. M. (2010). Bedtools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6):841–842.

[R Development Core Team, 2008] R Development Core Team (2008). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

[Ranger et al., 1998] Ranger, A. M., Grusby, M. J., Hodge, M. R., Gravallese, E. M., de la Brousse, F. C., Hoey, T., Mickanin, C., Baldwin, H. S., and Glimcher, L. H. (1998). The transcription factor NF-ATc is essential for cardiac valve formation. *Nature*, 392(6672):186–190.

[Reik and Walter, 2001] Reik, W. and Walter, J. (2001). Genomic imprinting: parental influence on the genome. *Nature Review Genetics*, 2(1):21–32.

[Robson et al., 2001] Robson, P., Stein, P., Zhou, B., Schultz, R. M., and Baldwin, H. (2001). Inner cell mass-specific expression of a cell adhesion molecule (PECAM-1/CD31) in the mouse blastocyst. *Developmental Biology*, 234(2):317 – 329.

[Rougier et al., 1998] Rougier, N., Bourc'his, D., Gomes, D. M., Niveleau, A., Plachot, M., Pàldi, A., and Viegas-Péquignot, E. (1998). Chromosome methylation patterns during mammalian preimplantationdevelopment. *Genes & Development Development*, 12(14):2108–2113.

[Sakamoto et al., 2007] Sakamoto, Y., Hara, K., Kanai-Azuma, M., Matsui, T., Miura, Y., Tsunekawa, N., Kurohmaru, M., Saijoh, Y., Koopman, P., and Kanai, Y. (2007). Redundant roles of Sox17 and Sox18 in early cardiovascular development of mouse embryos. *Biochemical and Biophysical Research Communications*, 360(3):539 – 544.

[Schachterle et al., 2012] Schachterle, W., Rojas, A., Xu, S.-M., and Black, B. L. (2012). ETS-dependent regulation of a distal Gata4 cardiac enhancer. *Developmental Biology*, 361(2):439 – 449.

[Schmidt et al., 2010] Schmidt, D., Schwalie, P. C., Ross-Innes, C. S., Hurtado, A., Brown, G. D., Carroll, J. S., Flicek, P., and Odom, D. T. (2010). A ctcf-independent role for cohesin in tissue-specific transcription. *Genome Research*, 20(5):578–588.

[Schmidt et al., 2012] Schmidt, D., Schwalie, P. C., Wilson, M. D., Ballester, B., Goncalves, A., Kutter, C., Brown, G. D., Marshall, A., Flicek, P., and Odom, D. T. (2012). Waves of retrotransposon expansion remodel genome organization and CTCF binding in multiple mammalian lineages. *Cell*, 148(1-2):335 – 348.

[Schulz et al., 2008] Schulz, R., Woodfine, K., Menheniott, T. R., Bourc'his, D., Bestor, T., and Oakey, R. J. (2008). WAMIDEX: A web atlas of murine genomic imprinting and differential expression. *Epigenetics*, 3(2):89–96.

[Schumacher et al., 2013] Schumacher, J. A., Bloomekatz, J., Garavito-Aguilar, Z. V., and Yelon, D. (2013). Tal1 regulates the formation of intercellular junctions and the maintenance of identity in the endocardium. *Developmental Biology*, 383(2):214 – 226.

[Seisenberger et al., 2012] Seisenberger, S., Andrews, S., Krueger, F., Arand, J., Walter, J., Santos, F., Popp, C., Thienpont, B., Dean, W., and Reik, W. (2012). The dynam-

ics of genome-wide DNA methylation reprogramming in mouse primordial germ cells. *Molecular Cell*, 48(6):849 – 862.

[Seitan et al., 2013] Seitan, V. C., Faure, A. J., Zhan, Y., McCord, R. P., Lajoie, B. R., Ing-Simmons, E., Lenhard, B., Giorgetti, L., Heard, E., Fisher, A. G., Flicek, P., Dekker, J., and Merkenschlager, M. (2013). Cohesin-based chromatin interactions enable regulated gene expression within preexisting architectural compartments. *Genome Research*, 23(12):2066–2077.

[Sharif et al., 2007] Sharif, J., Muto, M., Takebayashi, S.-i., Suetake, I., Iwamatsu, A., Endo, T. A., Shinga, J., Mizutani-Koseki, Y., Toyoda, T., Okamura, K., Tajima, S., Mitsuya, K., Okano, M., and Koseki, H. (2007). The SRA protein Np95 mediates epigenetic inheritance by recruiting Dnmt1 to methylated DNA. *Nature*, 450(7171):908–912.

[Shen et al., 2009] Shen, H.-L., Xu, Z.-G., Huang, L.-Y., Liu, D., Lin, D.-H., Cao, J.-B., Zhang, X., Wang, Z.-Q., Wang, W.-H., Yang, P.-Y., and Han, Z.-G. (2009). Liver-specific ZP domain-containing protein (LZP) as a new partner of Tamm-Horsfall protein harbors on renal tubules. *Molecular and Cellular Biochemistry*, 321(1-2):73–83.

[Shin et al., 2009] Shin, H., Liu, T., Manrai, A. K., and Liu, X. S. (2009). CEAS: cis-regulatory element annotation system. *Bioinformatics*, 25(19):2605–2606.

[Shipony et al., 2014] Shipony, Z., Mukamel, Z., Cohen, N. M., Landan, G., Chomsky, E., Zeliger, S. R., Fried, Y. C., Ainbinder, E., Friedman, N., and Tanay, A. (2014). Dynamic and static maintenance of epigenetic memory in pluripotent and somatic cells. *Nature*, 513(7516):115–119.

[Skibbens et al., 2013] Skibbens, R. V., Colquhoun, J. M., Green, M. J., Molnar, C. A., Sin, D. N., Sullivan, B. J., and Tanzosh, E. E. (2013). Cohesinopathies of a feather flock together. *PLoS Genet*, 9(12):e1004036.

[Slack, 2002] Slack, J. M. W. (2002). Conrad hal waddington: the last renaissance biologist? *Nat Rev Genet*, 3(11):889–895.

[Smith, 2001] Smith, A. G. (2001). Embryo-derived stem cells: Of mice and men. *Annual Review of Cell and Developmental Biology*, 17(1):435–462.

[Smith and Meissner, 2013] Smith, Z. D. and Meissner, A. (2013). DNA methylation: roles in mammalian development. *Nat Rev Genet*, 14(3):204–220.

[Smyth, 2005] Smyth, G. K. (2005). Limma: linear models for microarray data. In Gentleman, R., Carey, V., Dudoit, S., Irizarry, R., and Huber, W., editors, *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*, pages 397–420. Springer, New York.

[Sofueva et al., 2013] Sofueva, S., Yaffe, E., Chan, W.-C., Georgopoulou, D., Vietri Rudan, M., Mira-Bontenbal, H., Pollard, S. M., Schroth, G. P., Tanay, A., and Hadjur, S. (2013). Cohesin-mediated interactions organize chromosomal domain architecture. *The EMBO Journal*, 32(24):3119–3129.

[Stedman et al., 2008] Stedman, W., Kang, H., Lin, S., Kissil, J. L., Bartolomei, M. S., and Lieberman, P. M. (2008). Cohesins localize with CTCF at the KSHV latency control region and at cellular c-myc and H19/Igf2 insulators. *The EMBO Journal*, 27(4):654–666.

[Surani et al., 1984] Surani, M. A. H., Barton, S. C., and Norris, M. L. (1984). Development of reconstituted mouse eggs suggests imprinting of the genome during gametogenesis. *Nature*, 308(5959):548–550.

[Takahashi et al., 2014] Takahashi, R., Nagayama, S., Furu, M., Kajita, Y., Jin, Y., Kato, T., Imoto, S., Sakai, Y., and Toguchida, J. (2014). AFAP1L1, a novel associating partner with vinculin, modulates cellular morphology and motility, and promotes the progression of colorectal cancers. *Cancer Medicine*.

[Tam and Loebel, 2007] Tam, P. P. L. and Loebel, D. A. F. (2007). Gene function in mouse embryogenesis: get set for gastrulation. *Nat Rev Genet*, 8(5):368–381.

[The ENCODE Project Consortium, 2012] The ENCODE Project Consortium (2012). An integrated encyclopedia of dna elements in the human genome. *Nature*, 489(7414):57–74.

[Tian et al., 2014] Tian, X., Hu, T., Zhang, H., He, L., Huang, X., Liu, Q., Yu, W., He, L., Yang, Z., Yan, Y., Yang, X., Zhong, T. P., Pu, W. T., and Zhou, B. (2014). De novo formation of a distinct coronary vascular population in neonatal heart. *Science*, 345(6192):90–94.

[Tingare et al., 2013] Tingare, A., Thienpont, B., and Roderick, H. (2013). Epigenetics in the heart: the role of histone modifications in cardiac remodelling. *Biochemical Society transactions*, 41(3):789–796.

[Trapnell et al., 2013] Trapnell, C., Hendrickson, D. G., Sauvageau, M., Goff, L., Rinn, J. L., and Pachter, L. (2013). Differential analysis of gene regulation at transcript resolution with RNA-Seq. *Nature Biotechnology*, 31(1):46–53.

[Trapnell et al., 2010] Trapnell, C., Williams, B. A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M. J., Salzberg, S. L., Wold, B. J., and Pachter, L. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature Biotechnology*, 28(5):511–515.

[Treutlein et al., 2014] Treutlein, B., Brownfield, D. G., Wu, A. R., Neff, N. F., Mantalas, G. L., Espinoza, F. H., Desai, T. J., Krasnow, M. A., and Quake, S. R. (2014). Reconstructing lineage hierarchies of the distal lung epithelium using single-cell RNA-seq. *Nature*, 509(7500):371–375.

[Tutarel, 2014] Tutarel, O. (2014). Acquired heart conditions in adults with congenital heart disease: a growing problem. *Heart*, 100(17):1317–1321.

[Vallaster et al., 2012] Vallaster, M., Vallaster, C. D., and Wu, S. M. (2012). Epigenetic mechanisms in cardiac development and disease. *Acta Biochimica et Biophysica Sinica*, 44(1):92–102.

[Valouev et al., 2008] Valouev, A., Johnson, D. S., Sundquist, A., Medina, C., Anton, E., Batzoglou, S., Myers, R. M., and Sidow, A. (2008). Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data. *Nature Methods*, 5(9):829–834.

[Van Handel et al., 2012] Van Handel, B., Montel-Hagen, A., Sasidharan, R., Nakano, H., Ferrari, R., Boogerd, C. J., Schredelseker, J., Wang, Y., Hunter, S., Org, T., Zhou, J., Li, X., Pellegrini, M., Chen, J.-N., Orkin, S. H., Kurdistani, S. K., Evans, S. M., Nakano, A., and Mikkola, H. K. (2012). Scl represses cardiomyogenesis in prospective hemogenic endothelium and endocardium. *Cell*, 150(3):590 – 605.

[Vijayaraj et al., 2012] Vijayaraj, P., Le Bras, A., Mitchell, N., Kondo, M., Juliao, S., Wasserman, M., Beeler, D., Spokes, K., Aird, W. C., Baldwin, H. S., and Oettgen, P.

(2012). Erg is a crucial regulator of endocardial-mesenchymal transformation during cardiac valve morphogenesis. *Development*, 139(21):3973–3985.

[von Gise and Pu, 2012] von Gise, A. and Pu, W. T. (2012). Endocardial and epicardial epithelial to mesenchymal transitions in heart development and disease. *Circulation Research*, 110(12):1628–1645.

[Vostrov and Quitschke, 1997] Vostrov, A. A. and Quitschke, W. W. (1997). The zinc finger protein CTCF binds to the APBβ domain of the amyloid β-protein precursor promoter: Evidence for a role in transcriptional activation. *Journal of Biological Chemistry*, 272(52):33353–33359.

[Waddington, 2012] Waddington, C. H. (2012). The epigenotype. *International Journal of Epidemiology*, 41(1):10–13.

[Wamstad et al., 2012] Wamstad, J., Alexander, J., Truty, R., Shrikumar, A., Li, F., Eilertson, K., Ding, H., Wylie, J., Pico, A., Capra, J., Erwin, G., Kattman, S., Keller, G., Srivastava, D., Levine, S., Pollard, K., Holloway, A., Boyer, L., and Bruneau, B. (2012). Dynamic and coordinated epigenetic regulation of developmental transitions in the cardiac lineage. *Cell*, 151(1):206–220.

[Wang et al., 2009] Wang, Z., Gerstein, M., and Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nature Review Genetics*, 10(1):57–63.

[Wansleeben et al., 2011] Wansleeben, C., van Gurp, L., Feitsma, H., Kroon, C., Rieter, E., Verberne, M., Guryev, V., Cuppen, E., and Meijlink, F. (2011). An ENU-mutagenesis screen in the mouse: Identification of novel developmental gene functions. *PLoS ONE*, 6(4):e19357.

[Wat and Wat, 2014] Wat, J. J. and Wat, M. J. (2014). Sox7 in vascular development: review, insights and potential mechanisms. *The International Journal of Developmental Biology*, 58(1):1–8.

[Wei et al., 2010] Wei, G.-H., Badis, G., Berger, M. F., Kivioja, T., Palin, K., Enge, M., Bonke, M., Jolma, A., Varjosalo, M., Gehrke, A. R., Yan, J., Talukder, S., Turunen, M., Taipale, M., Stunnenberg, H. G., Ukkonen, E., Hughes, T. R., Bulyk, M. L., and Taipale, J. (2010). Genome-wide analysis of ETS-family DNA-binding in vitro and in vivo. *The EMBO Journal*, 29(13):2147–2160.

[Weirauch et al., 2014] Weirauch, M. T., Yang, A., Albu, M., Cote, A. G., Montenegro-Montero, A., Drewe, P., Najafabadi, H. S., Lambert, S. A., Mann, I., Cook, K., Zheng, H., Goity, A., van Bakel, H., Lozano, J.-C., Galli, M., Lewsey, M. G., Huang, E., Mukherjee, T., Chen, X., Reece-Hoyes, J. S., Govindarajan, S., Shaulsky, G., Walhout, A. J. M., Bouget, F.-Y., Ratsch, G., Larrondo, L. F., Ecker, J. R., and Hughes, T. R. (2014). Determination and inference of eukaryotic transcription factor sequence specificity. *Cell*, 158(6):1431–1443.

[Wendt et al., 2008] Wendt, K. S., Yoshida, K., Itoh, T., Bando, M., Koch, B., Schirghuber, E., Tsutsumi, S., Nagae, G., Ishihara, K., Mishiro, T., Yahata, K., Imamoto, F., Aburatani, H., Nakao, M., Imamoto, N., Maeshima, K., Shirahige, K., and Peters, J.-M. (2008). Cohesin mediates transcriptional insulation by CCCTC-binding factor. *Nature*, 451(7180):796–801.

[Wilbanks and Facciotti, 2010] Wilbanks, E. G. and Facciotti, M. T. (2010). Evaluation of algorithm performance in ChIP-Seq peak detection. *PLoS ONE*, 5(7):e11471.

[Wilkins, 2008] Wilkins, J. F. (2008). *Genomic imprinting*. Springer.

[Williamson et al., 2014] Williamson, C., Blake, A., Thomas, S., Beechey, C., Hancock, J., Cattanach, B., and Peters, J. (2014). World wide web site - mouse imprinting data and references. *unpublished*. http://www.har.mrc.ac.uk/research/genomic_imprinting/.

[Winderlich et al., 2009] Winderlich, M., Keller, L., Cagna, G., Broermann, A., Kamenyeva, O., Kiefer, F., Deutsch, U., Nottebaum, A. F., and Vestweber, D. (2009). VE-PTP controls blood vessel development by balancing Tie-2 activity. *The Journal of Cell Biology*, 185(4):657–671.

[Wolf and Brandvain, 2014] Wolf, J. B. and Brandvain, Y. (2014). Gene interactions in the evolution of genomic imprinting. *Heredity*, 113(2):129–137.

[Wolf and Hager, 2006] Wolf, J. B. and Hager, R. (2006). A maternal–offspring coadaptation theory for the evolution of genomic imprinting. *PLoS Biol*, 4(12):e380.

[Wolf and Hager, 2009] Wolf, J. B. and Hager, R. (2009). Selective abortion and the evolution of genomic imprinting. *Journal of Evolutionary Biology*, 22(12):2519–2523.

[Wossidlo et al., 2011] Wossidlo, M., Nakamura, T., Lepikhov, K., Marques, C. J., Zakhartchenko, V., Boiani, M., Arand, J., Nakano, T., Reik, W., and Walter, J. (2011).

5-Hydroxymethylcytosine in the mammalian zygote is linked with epigenetic reprogramming. *Nat Commun*, 2:241.

[Wu et al., 2013] Wu, B., Baldwin, H. S., and Zhou, B. (2013). Nfatc1 directs the endocardial progenitor cells to make heart valve primordium. *Trends in Cardiovascular Medicine*, 23(8):294–300.

[Wu et al., 2011a] Wu, B., Wang, Y., Lui, W., Langworthy, M., Tompkins, K. L., Hatzopoulos, A. K., Baldwin, H. S., and Zhou, B. (2011a). Nfatc1 coordinates valve endocardial cell lineage development required for heart valve formation. *Circulation Research*, 109(2):183–192.

[Wu et al., 2012] Wu, B., Zhang, Z., Lui, W., Chen, X., Wang, Y., Chamberlain, A. A., Moreno-Rodriguez, R., Markwald, R., ORourke, B., Sharp, D., Zheng, D., Lenz, J., Baldwin, H. ., Chang, C.-P., and Zhou, B. (2012). Endocardial cells form the coronary arteries by angiogenesis through myocardial-endocardial VEGF signaling. *Cell*, 151(5):1083–1096.

[Wu et al., 2011b] Wu, G., Yi, N., Absher, D., and Zhi, D. (2011b). Statistical quantification of methylation levels by next-generation sequencing. *PLoS ONE*, 6(6):e21034.

[Wu et al., 2005] Wu, J., Iwata, F., Grass, J. A., Osborne, C. S., Elnitski, L., Fraser, P., Ohneda, O., Yamamoto, M., and Bresnick, E. H. (2005). Molecular determinants of notch4 transcription in vascular endothelium. *Molecular and Cellular Biology*, 25(4):1458–1474.

[Wythe et al., 2013] Wythe, J. D., Dang, L. T., Devine, W. P., Boudreau, E., Artap, S. T., He, D., Schachterle, W., Stainier, D. Y., Oettgen, P., Black, B. L., Bruneau, B. G., and Fish, J. E. (2013). ETS factors regulate vegf-dependent arterial specification. *Developmental Cell*, 26(1):45 – 58.

[Xie et al., 2012] Xie, W., Barr, C. L., Kim, A., Yue, F., Lee, A. Y., Eubanks, J., Dempster, E. L., and Ren, B. (2012). Base-resolution analyses of sequence and parent-of-origin dependent DNA methylation in the mouse genome. *Cell*, 148(4):816 – 831.

[Xie et al., 2007] Xie, X., Mikkelsen, T. S., Gnirke, A., Lindblad-Toh, K., Kellis, M., and Lander, E. S. (2007). Systematic discovery of regulatory motifs in conserved regions

of the human genome, including thousands of ctcf insulator sites. *Proceedings of the National Academy of Sciences of the United States of America*, 104(17):7145–7150.

[Yalcin et al., 2012] Yalcin, B., Adams, D., Flint, J., and Keane, T. (2012). Next-generation sequencing of experimental mouse strains. *Mammalian genome : official journal of the International Mammalian Genome Society*, 23(9-10):490—498.

[Yan et al., 2012] Yan, B., Zhang, Z.-Z., Huang, L.-Y., Shen, H.-L., and Han, Z.-G. (2012). OIT3 deficiency impairs uric acid reabsorption in renal tubule. *FEBS Letters*, 586(6):760 – 765.

[Yan et al., 2010] Yan, M. S.-C., Matouk, C. C., and Marsden, P. A. (2010). Epigenetics of the vascular endothelium. *Journal of Applied Physiology*, 109(3):916–926.

[Yoon et al., 2007] Yoon, Y. S., Jeong, S., Rong, Q., Park, K.-Y., Chung, J. H., and Pfeifer, K. (2007). Analysis of the H19ICR insulator. *Molecular and Cellular Biology*, 27(9):3499–3510.

[Yu, 2014] Yu, G. (2014). *ChIPseeker: ChIPseeker for ChIP peak Annotation, Comparison, and Visualization*. R package version 1.2.3.

[Zabidi et al., 2014] Zabidi, M. A., Arnold, C. D., Schernhuber, K., Pagani, M., Rath, M., Frank, O., and Stark, A. (2014). Enhancer–core-promoter specificity separates developmental and housekeeping gene regulation. *Nature*, 518(7540):556–559.

[Zambon et al., 2012] Zambon, A. C., Gaj, S., Ho, I., Hanspers, K., Vranizan, K., Evelo, C. T., Conklin, B. R., Pico, A. R., and Salomonis, N. (2012). GO-Elite: a flexible solution for pathway and ontology over-representation. *Bioinformatics*, 28(16):2209–2210.

[Zhang et al., 2012] Zhang, H.-M., Chen, H., Liu, W., Liu, H., Gong, J., Wang, H., and Guo, A.-Y. (2012). Animaltfdb: a comprehensive animal transcription factor database. *Nucleic Acids Research*, 40(D1):D144–D149.

[Zhang et al., 2008] Zhang, Y., Liu, T., Meyer, C., Eeckhoute, J., Johnson, D., Bernstein, B., Nusbaum, C., Myers, R., Brown, M., Li, W., and Liu, X. S. (2008). Model-based analysis of chip-seq (macs). *Genome Biology*, 9(9):R137.

[Zheng, 2014] Zheng, X. (2014). Myocardin and smooth muscle differentiation. *Archives of Biochemistry and Biophysics*, 543(0):48 – 56.

[Zhou et al., 2005] Zhou, B., Wu, B., Tompkins, K. L., Boyer, K. L., Grindley, J. C., and Baldwin, H. S. (2005). Characterization of nfatc1 regulation identifies an enhancer required for gene expression that is specific to pro-valve endocardial cells in the developing heart. *Development*, 132(5):1137–1146.

[Zhu et al., 2013] Zhu, L. J., Pages, H., Gazin, C., Lawson, N., Ou, J., Lin, S., Lapointe, D., and Green, M. (2013). *ChIPpeakAnno: Batch annotation of the peaks identified from either ChIP-seq, ChIP-chip experiments or any experiments resulted in large number of chromosome ranges.* R package version 2.10.0.

[Zilbauer et al., 2013] Zilbauer, M., Rayner, T. F., Clark, C., Coffey, A. J., Joyce, C. J., Palta, P., Palotie, A., Lyons, P. A., and Smith, K. G. C. (2013). Genome-wide methylation analyses of primary human leukocyte subsets identifies functionally important cell-type–specific hypomethylated regions. *Blood*, 122(25):e52–e60.

Appendix A

Supplementary Data

A.1 Primer Sequences

Table A.1: Primer Sequences for *Magel2* Promoter Methylation Analysis

Primer Identifier	Sequence
SP6	TATTTAGGTGACACT
T7	TAATACGACTCACTATAGGG
BSMagel F1	GTGTTTGTTGAGAGTTGTTGAGAGA
BSMagel R1	ACCAAACAACCATAAAAACCTACAA

A.2 Differentially Methylated CGIs between the Endocardium and Endothelium

Table A.2: List of first 100 differentially methylated CGIs between the Endocardium and the Endothelium ordered by increasing p-value.

Locus	Endocardial Methylation	Endothelial Methylation	Methylation Difference	p-value	FDR
chr11:96849230-96849675	61.7	19.3	42.4	8.42E-03	0.35
chr15:25293805-25294929	62.8	21.3	41.5	9.03E-03	0.35
chr10:79858041-79858366	85.1	42.1	43.1	1.02E-02	0.35
chr7:53604045-53604704	54.9	16.3	38.6	1.03E-02	0.35
chr18:25636808-25637256	79.4	36.6	42.8	1.04E-02	0.35
chr7:126883244-126883727	53.3	15.2	38.1	1.05E-02	0.35
chr2:17821267-17821621	86.5	44.5	42.0	1.06E-02	0.35
chr11:104432676-104433180	66.1	27.1	39.0	1.09E-02	0.35
chr10:62884306-62884659	90.0	51.0	39.0	1.10E-02	0.35
chr12:118397878-118398412	59.3	21.7	37.6	1.14E-02	0.35
chr7:50898665-50899035	87.0	48.0	39.0	1.22E-02	0.35
chr13:59940407-59940888	85.8	46.3	39.5	1.22E-02	0.35
chr16:21332922-21333753	37.2	6.8	30.4	1.24E-02	0.35
chr8:87519536-87520140	80.0	40.0	40.0	1.24E-02	0.35
chr10:14479796-14480502	24.7	2.0	22.7	1.24E-02	0.35
chr7:108209677-108210318	49.8	15.0	34.8	1.26E-02	0.35
chr17:27856866-27857115	78.3	38.9	39.4	1.27E-02	0.35
chr7:20011351-20011713	79.3	40.0	39.2	1.29E-02	0.35
chr14:22505660-22506378	68.6	31.7	36.8	1.30E-02	0.35
chrX:71166644-71167703	20.1	0.7	19.5	1.32E-02	0.35
chr7:50930311-50931282	79.3	40.5	38.9	1.32E-02	0.35
chr4:114321126-114321723	58.8	23.6	35.3	1.33E-02	0.35
chr12:81066267-81066916	56.0	21.3	34.8	1.35E-02	0.35
chr10:79599013-79599571	72.7	35.8	37.0	1.37E-02	0.35
chrX:138740031-138740707	60.4	25.4	35.0	1.37E-02	0.35
chr9:109954396-109954859	76.4	38.7	37.7	1.38E-02	0.35
chr14:99712200-99712722	72.6	36.2	36.3	1.42E-02	0.35
chr2:29700771-29702248	46.1	14.0	32.1	1.45E-02	0.35
chr11:5162974-5163465	78.8	41.6	37.2	1.46E-02	0.35
chr10:93774282-93775331	27.3	4.5	22.8	1.48E-02	0.35
chr16:13867911-13868628	56.4	22.9	33.5	1.48E-02	0.35
chr12:75050520-75051192	49.5	17.0	32.5	1.49E-02	0.35
chr7:135010165-135010795	90.4	58.3	32.1	1.51E-02	0.35
chr4:124583319-124583859	24.2	3.4	20.8	1.51E-02	0.35
chr16:14292328-14292801	62.4	28.5	33.8	1.51E-02	0.35
chr16:33517179-33517756	43.7	12.9	30.8	1.52E-02	0.35
chr12:84961715-84962239	77.7	41.3	36.4	1.52E-02	0.35
chr3:65197122-65197496	87.0	52.6	34.4	1.53E-02	0.35
chr9:63801197-63802016	79.5	43.1	36.4	1.54E-02	0.35
chr1:75356650-75358015	45.7	14.5	31.2	1.54E-02	0.35

chr2:173935288-173936859	41.4	11.9	29.5	1.57E-02	0.35
chr3:95044597-95045098	70.5	36.0	34.4	1.58E-02	0.35
chr14:106292368-106292828	47.3	16.0	31.3	1.58E-02	0.35
chr17:34048518-34048958	82.6	46.9	35.7	1.58E-02	0.35
chr6:84055242-84056106	32.3	7.3	25.0	1.59E-02	0.35
chr15:102780056-102780582	52.3	20.2	32.1	1.59E-02	0.35
chr1:180298620-180299510	23.8	3.7	20.1	1.61E-02	0.35
chr8:109632921-109633464	79.8	44.2	35.6	1.62E-02	0.35
chr1:182824345-182825199	54.9	23.1	31.9	1.65E-02	0.35
chr8:124217435-124217840	77.9	42.9	35.0	1.66E-02	0.35
chr11:99793315-99793744	78.0	43.0	35.0	1.66E-02	0.35
chr4:150462668-150463188	86.4	53.2	33.1	1.67E-02	0.35
chr5:61459925-61460401	96.7	70.9	25.8	1.68E-02	0.35
chr13:113441941-113442916	82.2	47.5	34.7	1.68E-02	0.35
chr5:116003836-116004208	87.9	55.9	31.9	1.68E-02	0.35
chr10:5977292-5977533	84.2	50.2	34.0	1.69E-02	0.35
chr15:25710655-25711741	82.3	47.9	34.4	1.71E-02	0.35
chr5:97516050-97516941	88.9	58.3	30.7	1.72E-02	0.35
chr4:139560413-139560892	66.2	33.8	32.4	1.74E-02	0.35
chr7:134358010-134358521	72.7	39.2	33.4	1.74E-02	0.35
chr9:54539061-54539756	48.2	17.9	30.2	1.74E-02	0.35
chr9:27598695-27599548	32.5	8.4	24.2	1.76E-02	0.35
chr2:59681213-59681733	88.8	58.4	30.4	1.76E-02	0.35
chr8:123407141-123407674	65.8	33.7	32.1	1.77E-02	0.35
chr10:20119711-20120527	83.0	49.4	33.6	1.77E-02	0.35
chr19:46781729-46782218	48.2	18.2	30.0	1.78E-02	0.35
chr3:89103529-89104014	51.9	21.3	30.6	1.78E-02	0.35
chr5:143690518-143691443	20.8	3.2	17.6	1.78E-02	0.35
chr3:159158100-159158671	13.8	0.4	13.4	1.79E-02	0.35
chr18:61808674-61809085	86.0	54.2	31.9	1.80E-02	0.35
chr17:45941983-45943077	40.9	13.2	27.7	1.80E-02	0.35
chr8:86296406-86297097	48.6	18.7	29.9	1.80E-02	0.35
chr7:31640030-31640428	89.4	60.0	29.4	1.81E-02	0.35
chr4:115796125-115796544	12.5	0.0	12.5	1.84E-02	0.35
chr5:149842687-149843490	50.0	20.2	29.8	1.84E-02	0.35
chr1:44255391-44256133	75.4	42.5	33.0	1.85E-02	0.35
chr8:86184960-86185394	74.4	41.6	32.8	1.85E-02	0.35
chr13:93564256-93565180	49.7	20.0	29.7	1.86E-02	0.35
chr4:43419193-43419680	20.6	3.4	17.2	1.86E-02	0.35
chr5:120141042-120141437	14.2	0.8	13.4	1.86E-02	0.35
chr6:112896704-112897438	35.8	10.7	25.1	1.87E-02	0.35
chr2:151401457-151402277	27.4	6.5	20.9	1.87E-02	0.35
chr10:81816853-81817518	16.1	1.6	14.5	1.87E-02	0.35
chr17:71011650-71012339	89.2	60.3	28.8	1.88E-02	0.35

chr8:109241777-109242386	63.1	32.3	30.9	1.90E-02	0.35
chr17:13102514-13103073	77.9	45.1	32.8	1.90E-02	0.35
chr6:47974378-47975698	75.6	43.1	32.5	1.91E-02	0.35
chr11:97361128-97362869	61.6	31.0	30.6	1.92E-02	0.35
chr14:76654766-76655262	61.2	30.7	30.5	1.92E-02	0.35
chr1:72908839-72909278	89.5	61.4	28.1	1.93E-02	0.35
chr3:88011040-88011606	31.2	8.6	22.6	1.93E-02	0.35
chr15:12763065-12764056	82.7	50.6	32.1	1.93E-02	0.35
chr6:89185575-89186168	84.5	53.3	31.2	1.95E-02	0.35
chr11:94253874-94254371	56.3	26.4	29.9	1.95E-02	0.35
chr4:53488083-53488470	82.6	50.7	31.9	1.96E-02	0.35
chr15:89377737-89379428	70.7	39.3	31.4	1.96E-02	0.35
chr4:138894006-138894469	78.6	46.4	32.2	1.98E-02	0.35
chr7:29329020-29329804	54.9	25.3	29.5	1.98E-02	0.35
chr1:89371543-89372426	36.2	11.5	24.7	1.98E-02	0.35

A.3 Differentially Methylated Genomic Regions between the Endocardium and the Endothelium

Table A.3: Full list of differentially methylated genomic loci between the endocardium and the endothelium as identified by the analysis utilising BSmooth, ordered by methylation difference between the two tissues.

Genomic Locus	Area T-statistic	Endocardial Methylation	Endothelial Methylation	Mean Methylation Difference	Methylation in Endocardium
chr11:20550017-20550548	121.61	81.1%	30.2%	50.9%	hyper
chrX:154006656-154007063	123.92	86.8%	36.9%	50.0%	hyper
chr11:96817033-96817576	164.59	84.3%	35.5%	48.8%	hyper
chr4:153880639-153881227	113.16	73.4%	25.3%	48.1%	hyper
chr8:91769076-91769532	132.12	68.8%	23.9%	44.8%	hyper
chr18:25636890-25637087	108.62	79.8%	36.1%	43.7%	hyper
chr8:64028283-64028744	200.36	77.9%	36.5%	41.4%	hyper
chr15:25293946-25294414	309.43	67.5%	26.3%	41.2%	hyper
chr3:148585565-148586109	116.54	88.1%	47.8%	40.3%	hyper
chr7:50930490-50930781	81.40	80.2%	40.0%	40.2%	hyper
chr4:138894197-138894354	99.77	74.4%	34.3%	40.1%	hyper
chr4:117040019-117040677	147.29	84.2%	44.8%	39.4%	hyper
chr11:116967987-116968706	112.43	76.0%	36.6%	39.3%	hyper
chr5:7622306-7622586	76.25	81.2%	42.1%	39.1%	hyper
chr11:104432697-104433410	203.62	65.6%	26.6%	39.0%	hyper
chr9:63801417-63801870	196.96	80.0%	41.3%	38.6%	hyper
chr12:84961946-84962112	58.45	76.6%	38.7%	37.9%	hyper
chr10:26047327-26047995	116.69	79.1%	41.3%	37.8%	hyper
chr15:59482983-59483306	82.34	82.8%	45.0%	37.8%	hyper
chr11:114403263-114403975	116.45	77.9%	40.2%	37.8%	hyper
chr13:113442131-113442504	81.43	78.9%	41.4%	37.5%	hyper
chr14:55250863-55251098	82.27	82.6%	46.0%	36.7%	hyper
chr8:72817084-72817565	73.83	84.1%	47.6%	36.6%	hyper
chr4:123000050-123000422	177.37	66.1%	29.6%	36.4%	hyper
chr14:8972127-8973204	111.57	70.9%	34.5%	36.4%	hyper
chr7:53604302-53604654	101.23	53.7%	17.4%	36.3%	hyper
chr6:142845417-142846067	88.96	83.2%	47.0%	36.2%	hyper
chr11:97811095-97811915	177.03	78.3%	42.1%	36.1%	hyper
chr9:83557758-83558664	162.18	76.0%	40.0%	36.0%	hyper
chr12:119898955-119900124	75.74	62.0%	26.1%	35.9%	hyper
chr2:24534514-24534988	209.60	81.3%	45.5%	35.8%	hyper
chr9:110533954-110534504	95.38	65.1%	29.6%	35.5%	hyper
chr17:27856820-27857320	166.46	80.5%	45.1%	35.4%	hyper
chr18:46876036-46876695	107.88	82.0%	46.6%	35.4%	hyper
chr6:72680493-72680944	86.68	82.8%	47.7%	35.1%	hyper
chr4:125212061-125213486	130.58	50.7%	15.6%	35.0%	hyper
chr14:56519271-56519756	84.84	74.3%	39.3%	35.0%	hyper
chr11:28930542-28930822	98.69	80.7%	45.8%	34.9%	hyper
chr8:44876887-44877618	85.29	76.9%	42.1%	34.8%	hyper

chr2:31160427-31160688	65.00	81.4%	46.7%	34.7%	hyper
chr10:120752173-120752764	162.85	73.2%	38.5%	34.7%	hyper
chr6:52273000-52273700	100.62	63.5%	29.0%	34.6%	hyper
chr2:147702816-147703764	129.93	76.8%	42.3%	34.5%	hyper
chr10:20119828-20120526	263.17	83.3%	48.9%	34.4%	hyper
chr1:89178205-89179039	137.65	81.9%	47.5%	34.3%	hyper
chr7:13556679-13556924	87.58	82.2%	48.0%	34.2%	hyper
chr12:87720288-87721157	134.17	77.9%	43.7%	34.2%	hyper
chr7:133768119-133768628	89.56	83.7%	49.5%	34.2%	hyper
chr14:55098383-55098699	92.35	57.4%	23.2%	34.2%	hyper
chrX:138740315-138740713	107.79	58.0%	24.0%	34.1%	hyper
chr3:95044758-95045100	110.63	69.9%	35.9%	34.1%	hyper
chr1:179492014-179492920	167.67	89.1%	55.2%	33.9%	hyper
chr12:111736298-111736957	96.44	82.0%	48.1%	33.9%	hyper
chr15:27980776-27981645	226.39	69.1%	35.3%	33.8%	hyper
chr9:113897783-113898366	92.15	78.7%	44.9%	33.8%	hyper
chr2:168519698-168520288	102.38	64.9%	31.2%	33.7%	hyper
chr2:150237786-150238319	117.77	78.1%	44.4%	33.7%	hyper
chr5:144143301-144143635	97.99	79.6%	45.8%	33.7%	hyper
chr3:68385653-68386262	105.27	79.3%	45.7%	33.6%	hyper
chr5:151249330-151249927	81.29	84.4%	50.9%	33.6%	hyper
chr15:55708482-55708897	74.19	75.4%	42.0%	33.4%	hyper
chr10:120131616-120132067	86.54	71.8%	38.5%	33.4%	hyper
chr7:56887804-56888417	76.82	65.3%	32.0%	33.3%	hyper
chr7:149766621-149768046	287.28	87.8%	54.5%	33.3%	hyper
chr8:10805689-10806092	97.68	54.5%	21.3%	33.1%	hyper
chr14:29330909-29331627	135.31	72.3%	39.2%	33.1%	hyper
chr4:125222430-125222877	59.95	68.3%	35.2%	33.1%	hyper
chr3:21222765-21223114	88.76	84.7%	51.8%	33.0%	hyper
chr14:105826684-105827339	139.54	70.5%	37.5%	33.0%	hyper
chr15:27924612-27925488	113.33	69.3%	36.4%	32.9%	hyper
chr11:5163165-5163515	94.81	78.5%	45.6%	32.9%	hyper
chr9:114547118-114547781	76.87	81.9%	49.0%	32.9%	hyper
chr3:66900683-66900925	102.75	91.0%	58.1%	32.8%	hyper
chr8:109632819-109633508	121.87	78.5%	45.8%	32.8%	hyper
chr8:73269722-73269943	79.16	86.8%	54.0%	32.8%	hyper
chr10:79509265-79510337	72.25	69.9%	37.1%	32.7%	hyper
chr2:87366209-87366535	70.58	71.6%	38.8%	32.7%	hyper
chr4:136019237-136019486	69.33	46.7%	14.0%	32.7%	hyper
chr7:148211594-148212613	75.05	72.1%	39.5%	32.6%	hyper
chr10:93693157-93693587	111.40	77.0%	44.5%	32.5%	hyper
chr3:52760307-52760810	97.77	83.3%	50.8%	32.4%	hyper
chr5:65799006-65799736	126.34	74.9%	42.5%	32.4%	hyper
chr11:44645791-44646249	105.95	49.0%	16.6%	32.4%	hyper

chr11:120044927-120045408	96.34	66.8%	34.5%	32.4%	hyper
chr11:3387833-3388588	118.15	77.7%	45.3%	32.4%	hyper
chr12:12700495-12701526	244.17	57.8%	25.4%	32.3%	hyper
chr13:52975064-52976007	91.10	70.1%	37.9%	32.3%	hyper
chr7:78528053-78528603	118.63	70.4%	38.3%	32.2%	hyper
chr2:173191080-173191315	99.19	69.4%	37.3%	32.1%	hyper
chr3:153568685-153569035	155.32	73.1%	41.0%	32.1%	hyper
chr14:22336557-22336788	94.48	82.1%	50.0%	32.1%	hyper
chr9:96444280-96444757	77.76	71.2%	39.1%	32.0%	hyper
chr2:83532327-83533271	108.54	69.2%	37.2%	32.0%	hyper
chr15:25562473-25563160	84.10	87.2%	55.2%	31.9%	hyper
chr7:29957746-29958208	135.20	76.6%	44.7%	31.9%	hyper
chr4:114321291-114321618	106.51	57.3%	25.4%	31.9%	hyper
chr2:103843358-103843795	96.17	70.0%	38.1%	31.9%	hyper
chr4:126901908-126902545	115.12	60.1%	28.2%	31.9%	hyper
chr8:4206269-4206850	160.25	55.2%	23.4%	31.9%	hyper
chr6:23222610-23223675	130.42	51.8%	20.0%	31.8%	hyper
chr9:110817103-110817617	142.35	80.1%	48.3%	31.8%	hyper
chr11:76171083-76171334	65.22	73.8%	42.0%	31.8%	hyper
chr3:36962467-36962746	164.50	79.7%	47.9%	31.8%	hyper
chr1:129897370-129898094	71.72	85.8%	54.0%	31.8%	hyper
chr17:5418663-5419303	119.05	84.8%	53.0%	31.8%	hyper
chr5:33854489-33855593	131.17	79.7%	48.0%	31.7%	hyper
chr1:162962729-162963711	70.35	82.6%	51.0%	31.7%	hyper
chr8:73026052-73026374	62.30	61.9%	30.3%	31.6%	hyper
chr13:12199582-12200010	88.78	72.3%	40.7%	31.6%	hyper
chr17:18046692-18047335	245.33	69.4%	37.7%	31.6%	hyper
chr4:149033577-149034159	118.16	81.1%	49.5%	31.6%	hyper
chr18:69837437-69838038	125.39	86.4%	54.8%	31.6%	hyper
chr4:135846684-135847192	213.57	49.2%	17.6%	31.6%	hyper
chr8:86192552-86192929	99.51	69.2%	37.7%	31.5%	hyper
chr1:163373007-163373968	137.61	65.2%	33.7%	31.5%	hyper
chr10:58003391-58003733	121.81	80.8%	49.3%	31.4%	hyper
chr16:43142871-43143259	183.14	83.0%	51.6%	31.4%	hyper
chr4:14606912-14607206	72.75	78.5%	47.1%	31.4%	hyper
chr13:14057735-14058074	81.99	78.5%	47.2%	31.3%	hyper
chr9:39203548-39203848	77.02	81.9%	50.6%	31.3%	hyper
chr10:94638146-94638844	117.67	69.4%	38.2%	31.3%	hyper
chr12:85913238-85913949	82.72	58.7%	27.5%	31.2%	hyper
chr19:57064818-57065598	101.77	78.4%	47.2%	31.2%	hyper
chr8:46136332-46136765	131.64	87.5%	56.2%	31.2%	hyper
chr2:165658869-165659522	82.22	84.5%	53.3%	31.2%	hyper
chr2:106696495-106697317	86.39	74.6%	43.5%	31.1%	hyper
chr8:11279391-11279803	87.06	72.8%	41.8%	31.1%	hyper

chr11:115119008-115119217	70.91	66.2%	35.2%	31.0%	hyper
chr8:124859685-124859999	128.73	45.5%	14.5%	31.0%	hyper
chr12:118397824-118398305	108.47	58.0%	27.1%	31.0%	hyper
chr18:20759818-20760284	115.35	84.9%	53.9%	31.0%	hyper
chr15:93333574-93334094	145.65	83.1%	52.3%	30.8%	hyper
chr11:34964809-34965604	180.55	84.1%	53.3%	30.8%	hyper
chr12:106959841-106960413	76.92	79.6%	48.8%	30.8%	hyper
chr11:43466414-43466817	70.31	74.9%	44.1%	30.8%	hyper
chr17:45942669-45942994	128.33	38.0%	7.3%	30.7%	hyper
chrX:166419731-166420074	82.09	82.9%	52.3%	30.7%	hyper
chr1:154762425-154763580	65.91	80.1%	49.5%	30.7%	hyper
chr5:134714011-134714456	82.26	84.3%	53.6%	30.6%	hyper
chr11:96440246-96440681	113.92	66.7%	36.1%	30.6%	hyper
chr2:120965087-120965692	209.04	70.2%	39.6%	30.6%	hyper
chr15:8544799-8545602	96.68	82.5%	51.9%	30.6%	hyper
chr8:124217553-124217926	71.47	78.0%	47.4%	30.6%	hyper
chr15:55040178-55040661	106.41	78.1%	47.5%	30.6%	hyper
chr3:108187823-108188389	163.54	83.4%	52.8%	30.5%	hyper
chr2:29110478-29111264	114.35	76.2%	45.7%	30.5%	hyper
chr3:65460533-65461071	117.73	65.2%	34.7%	30.5%	hyper
chr9:109954369-109954889	124.45	75.4%	44.9%	30.5%	hyper
chr1:186634432-186634830	98.49	59.8%	29.4%	30.4%	hyper
chr5:114570042-114570523	66.19	84.0%	53.6%	30.4%	hyper
chr7:135010044-135010639	175.80	88.6%	58.2%	30.4%	hyper
chr3:50513551-50514177	86.49	78.7%	48.4%	30.4%	hyper
chr19:43647741-43648429	119.00	88.6%	58.3%	30.4%	hyper
chr17:69557035-69557497	89.65	80.7%	50.3%	30.3%	hyper
chr9:22707765-22708325	71.57	85.0%	54.7%	30.3%	hyper
chr7:127741433-127741781	116.41	61.3%	31.0%	30.3%	hyper
chr12:113958407-113959505	113.23	65.0%	34.8%	30.2%	hyper
chr11:116951603-116952944	95.73	78.4%	48.2%	30.2%	hyper
chr19:46798935-46799509	113.20	81.2%	51.0%	30.2%	hyper
chr8:98125047-98125535	80.22	80.0%	50.0%	30.1%	hyper
chr5:116003709-116004100	123.89	88.4%	58.3%	30.1%	hyper
chr10:19861332-19862037	77.87	80.3%	50.3%	30.1%	hyper
chr1:52546812-52547220	95.74	79.9%	49.8%	30.0%	hyper
chr2:26026217-26027130	89.19	67.7%	37.7%	30.0%	hyper
chr11:20556345-20557360	128.87	61.5%	31.5%	30.0%	hyper
chr11:106468008-106468818	77.68	81.0%	50.9%	30.0%	hyper
chr4:117003245-117004132	87.26	78.3%	48.3%	30.0%	hyper
chr8:23779930-23780438	83.60	87.8%	57.8%	30.0%	hyper
chr11:113570080-113570880	134.07	59.9%	30.0%	29.9%	hyper
chr12:88254354-88254948	96.80	86.3%	56.3%	29.9%	hyper
chr8:109241931-109242163	85.55	59.7%	29.7%	29.9%	hyper

chr13:99343432-99344453	266.12	80.0%	50.1%	29.9%	hyper
chr2:152253151-152253878	115.04	54.0%	24.1%	29.9%	hyper
chr4:26645556-26645945	78.76	78.3%	48.4%	29.9%	hyper
chr2:126186491-126187372	165.61	78.2%	48.4%	29.9%	hyper
chr7:138835717-138836274	105.72	77.1%	47.2%	29.9%	hyper
chr8:108090552-108090881	75.50	50.2%	20.4%	29.8%	hyper
chr5:103954986-103955263	65.17	90.0%	60.2%	29.8%	hyper
chr5:56235254-56235473	90.55	80.8%	51.0%	29.8%	hyper
chr8:73121547-73121732	65.54	51.1%	21.4%	29.7%	hyper
chrX:4906716-4907109	83.50	82.8%	53.1%	29.7%	hyper
chr4:101189001-101189294	81.81	90.7%	60.9%	29.7%	hyper
chr15:39945555-39946960	190.41	67.6%	37.9%	29.7%	hyper
chr9:96714563-96714689	95.84	89.4%	59.8%	29.7%	hyper
chr11:4139573-4139951	115.01	80.5%	50.9%	29.6%	hyper
chr15:102780161-102780416	80.93	53.0%	23.5%	29.6%	hyper
chr5:118887676-118888683	85.61	57.9%	28.4%	29.6%	hyper
chr11:85738410-85738910	84.01	73.6%	44.1%	29.6%	hyper
chr2:29736055-29736324	68.60	74.8%	45.3%	29.5%	hyper
chr2:17854469-17854935	72.99	64.4%	34.9%	29.5%	hyper
chr16:90132564-90133273	135.78	73.1%	43.7%	29.4%	hyper
chr2:18477581-18478236	95.65	80.9%	51.5%	29.4%	hyper
chr5:34035168-34035359	58.35	47.9%	18.5%	29.4%	hyper
chr6:148930872-148931889	77.66	79.3%	50.0%	29.4%	hyper
chr5:102317190-102317415	88.50	85.1%	55.8%	29.3%	hyper
chr6:134569347-134569912	91.37	86.3%	57.0%	29.3%	hyper
chr7:31299458-31300412	108.19	66.3%	37.0%	29.3%	hyper
chr7:119885156-119885680	89.87	66.5%	37.2%	29.3%	hyper
chr3:31107154-31108172	103.55	73.3%	44.0%	29.3%	hyper
chr15:100499800-100500090	63.25	85.6%	56.4%	29.3%	hyper
chr11:120122554-120122994	80.02	76.1%	46.9%	29.3%	hyper
chr12:74894209-74894870	118.88	61.8%	32.5%	29.3%	hyper
chr15:80780036-80781106	65.71	86.8%	57.5%	29.2%	hyper
chr9:56284989-56285562	179.15	86.6%	57.5%	29.1%	hyper
chr19:45641609-45642656	168.07	65.4%	36.3%	29.1%	hyper
chr7:20055181-20055460	96.30	73.0%	43.9%	29.1%	hyper
chr11:94011846-94012759	78.66	63.8%	34.7%	29.1%	hyper
chr17:32470744-32470984	74.77	86.3%	57.3%	29.1%	hyper
chr1:137114015-137114765	85.15	71.5%	42.4%	29.1%	hyper
chr11:51781864-51783034	99.54	84.0%	55.0%	29.0%	hyper
chr7:38264507-38264994	92.57	86.3%	57.4%	29.0%	hyper
chr4:124906417-124906697	80.62	77.6%	48.6%	29.0%	hyper
chr9:92347926-92348281	87.40	87.6%	58.7%	28.9%	hyper
chr7:144017249-144017637	92.09	78.9%	49.9%	28.9%	hyper
chr17:7408875-7409589	82.42	72.4%	43.5%	28.9%	hyper

chr5:53916450-53917327	141.88	81.7%	52.8%	28.9%	hyper
chr10:37984011-37984378	73.18	83.9%	55.0%	28.9%	hyper
chr2:162421258-162422269	96.39	86.6%	57.8%	28.9%	hyper
chr12:113027316-113027845	103.18	84.6%	55.8%	28.8%	hyper
chr4:136312041-136312761	91.77	61.9%	33.1%	28.8%	hyper
chr4:53487926-53488522	94.84	82.9%	54.1%	28.8%	hyper
chr8:113890877-113891293	73.39	75.8%	47.0%	28.8%	hyper
chr7:125633459-125634228	240.25	53.9%	25.1%	28.8%	hyper
chr11:29725201-29725537	128.03	74.1%	45.3%	28.8%	hyper
chr13:112538971-112539308	116.77	89.5%	60.7%	28.8%	hyper
chr10:26415113-26415689	114.85	55.2%	26.4%	28.8%	hyper
chr10:79616502-79616904	105.56	83.4%	54.6%	28.8%	hyper
chr8:82579585-82580748	128.78	71.3%	42.5%	28.8%	hyper
chr4:126002087-126002467	76.15	74.6%	45.8%	28.8%	hyper
chr8:107860607-107860809	59.26	73.6%	44.9%	28.7%	hyper
chr12:107167434-107168267	136.43	76.7%	48.0%	28.7%	hyper
chr6:125974444-125975093	95.28	86.9%	58.2%	28.7%	hyper
chr10:69902227-69902846	78.05	74.6%	45.9%	28.7%	hyper
chr6:50272626-50273168	117.61	84.2%	55.6%	28.6%	hyper
chr3:78870524-78870792	102.23	84.8%	56.2%	28.6%	hyper
chr2:118256446-118256940	155.81	80.9%	52.3%	28.6%	hyper
chr10:93986605-93987099	77.34	59.2%	30.6%	28.6%	hyper
chr9:46142581-46142954	98.72	83.4%	54.8%	28.6%	hyper
chr16:13130132-13130531	137.23	78.4%	49.8%	28.5%	hyper
chr7:119253915-119254613	95.30	85.5%	56.9%	28.5%	hyper
chr19:59207361-59208191	107.39	84.8%	56.3%	28.5%	hyper
chr15:72931701-72932223	93.84	87.2%	58.7%	28.5%	hyper
chr19:44060531-44061068	137.67	68.9%	40.4%	28.5%	hyper
chr11:117352477-117353581	82.68	83.7%	55.3%	28.5%	hyper
chr2:38375524-38376057	79.38	82.5%	54.0%	28.4%	hyper
chr4:154357692-154358075	82.48	68.2%	39.7%	28.4%	hyper
chr6:89185481-89185884	99.60	85.5%	57.0%	28.4%	hyper
chr13:57383559-57383795	71.76	72.0%	43.6%	28.4%	hyper
chr4:118108001-118108883	89.93	81.0%	52.6%	28.4%	hyper
chr19:7002201-7002565	66.49	82.9%	54.5%	28.3%	hyper
chr9:20442956-20443282	57.07	83.6%	55.3%	28.3%	hyper
chr9:106549597-106549960	132.66	81.2%	52.9%	28.3%	hyper
chr11:103703044-103703659	91.83	52.9%	24.6%	28.3%	hyper
chr7:107871216-107872491	87.71	78.2%	49.9%	28.3%	hyper
chr2:168098736-168099311	90.35	60.6%	32.3%	28.3%	hyper
chr10:122450555-122450964	107.01	75.3%	47.0%	28.3%	hyper
chr12:118685936-118686442	84.93	71.5%	43.2%	28.2%	hyper
chr12:13215767-13216338	88.71	82.1%	53.9%	28.2%	hyper
chr9:99699916-99700608	113.22	72.9%	44.7%	28.2%	hyper

chr19:57729872-57730414	94.27	83.6%	55.4%	28.2%	hyper
chr8:63070352-63070835	126.10	79.5%	51.3%	28.2%	hyper
chr11:101040115-101040325	61.00	80.7%	52.5%	28.2%	hyper
chr12:101419205-101419822	76.28	75.2%	47.0%	28.2%	hyper
chr5:150867030-150867522	89.23	83.0%	54.8%	28.2%	hyper
chr16:30657082-30657587	126.60	76.7%	48.6%	28.2%	hyper
chr18:75738997-75739687	86.72	78.4%	50.3%	28.2%	hyper
chr4:118190073-118190950	78.17	61.8%	33.6%	28.1%	hyper
chr10:94636961-94637753	77.42	84.3%	56.1%	28.1%	hyper
chr10:82360501-82360825	76.01	81.1%	53.0%	28.1%	hyper
chrX:107612618-107613043	62.16	79.4%	51.3%	28.1%	hyper
chr4:140580938-140581718	236.64	53.4%	25.3%	28.1%	hyper
chr14:121634025-121634409	105.29	85.3%	57.2%	28.1%	hyper
chr12:64086039-64086347	65.05	86.6%	58.6%	28.1%	hyper
chr5:73739585-73740019	84.93	71.8%	43.8%	28.1%	hyper
chr13:51503099-51503468	70.30	62.5%	34.4%	28.1%	hyper
chr15:61994205-61995363	78.77	78.6%	50.6%	28.0%	hyper
chr10:127095769-127096335	108.01	77.6%	49.6%	28.0%	hyper
chr5:99908196-99908938	91.57	64.1%	36.0%	28.0%	hyper
chr9:41184428-41184772	91.89	69.1%	41.1%	28.0%	hyper
chr3:34555551-34556010	64.44	41.7%	13.7%	28.0%	hyper
chr5:135442964-135443614	64.01	78.9%	50.9%	28.0%	hyper
chr5:92821616-92822293	146.27	80.1%	52.1%	27.9%	hyper
chr7:144444832-144445310	70.83	74.0%	46.1%	27.9%	hyper
chr6:88009696-88010378	74.74	79.7%	51.7%	27.9%	hyper
chr1:181540120-181540428	103.80	77.5%	49.6%	27.9%	hyper
chr15:97168983-97169893	106.90	59.0%	31.1%	27.9%	hyper
chr17:28685489-28685767	57.81	76.8%	48.9%	27.9%	hyper
chr2:38553060-38553686	103.68	84.5%	56.6%	27.9%	hyper
chr13:98517257-98517946	91.89	84.9%	57.0%	27.9%	hyper
chr3:90664657-90665000	179.74	79.1%	51.3%	27.8%	hyper
chr14:104471046-104471213	58.21	78.9%	51.1%	27.8%	hyper
chr11:102126990-102127590	125.74	82.1%	54.3%	27.8%	hyper
chr3:121787724-121788914	133.80	68.6%	40.8%	27.8%	hyper
chr3:84074010-84074594	112.06	63.4%	35.6%	27.8%	hyper
chr12:113299102-113299372	111.19	81.3%	53.5%	27.7%	hyper
chr9:43071934-43072567	83.40	70.4%	42.7%	27.7%	hyper
chr8:124793057-124793175	56.95	72.1%	44.4%	27.7%	hyper
chr9:104184599-104185389	107.93	71.8%	44.2%	27.7%	hyper
chr17:87777602-87777933	67.23	85.5%	57.9%	27.6%	hyper
chr4:139043171-139043699	58.71	83.2%	55.6%	27.6%	hyper
chr6:104015173-104015506	59.68	84.6%	57.0%	27.6%	hyper
chr1:136981566-136982234	97.23	58.0%	30.4%	27.5%	hyper
chr5:44490853-44491527	93.33	55.8%	28.3%	27.5%	hyper

chr4:62832135-62832636	153.21	80.9%	53.3%	27.5%	hyper
chr7:105313648-105314322	87.14	77.3%	49.8%	27.5%	hyper
chr5:122278315-122278607	143.71	62.6%	35.1%	27.5%	hyper
chr5:31354812-31355895	74.28	89.4%	61.8%	27.5%	hyper
chr5:113890672-113891443	154.88	73.1%	45.7%	27.5%	hyper
chr12:19249790-19250063	166.53	63.8%	36.4%	27.4%	hyper
chr16:22266672-22267571	111.09	78.7%	51.3%	27.4%	hyper
chr6:10919575-10920994	74.91	72.3%	44.9%	27.4%	hyper
chr5:27588319-27588751	85.72	47.3%	20.0%	27.3%	hyper
chr11:97777750-97778176	64.38	88.7%	61.4%	27.3%	hyper
chr4:138788360-138789119	66.50	58.4%	31.1%	27.3%	hyper
chr17:95050620-95051589	94.65	76.2%	48.9%	27.3%	hyper
chr10:58135347-58136460	121.53	69.0%	41.8%	27.3%	hyper
chr5:115894491-115895052	104.10	88.4%	61.2%	27.2%	hyper
chr7:126883460-126883769	90.51	55.2%	28.1%	27.2%	hyper
chr2:65656236-65656805	101.59	75.9%	48.8%	27.2%	hyper
chr12:70231282-70231733	82.44	65.8%	38.6%	27.1%	hyper
chr7:106620147-106620341	74.02	76.8%	49.7%	27.1%	hyper
chr2:165136190-165136557	93.54	79.0%	51.9%	27.1%	hyper
chr3:45382907-45383143	63.95	86.4%	59.3%	27.1%	hyper
chr8:88005832-88007157	95.02	75.4%	48.3%	27.1%	hyper
chr14:76654648-76655491	193.10	66.0%	38.9%	27.1%	hyper
chr1:182824490-182825286	184.10	54.9%	27.8%	27.1%	hyper
chr3:152398249-152398782	111.79	70.0%	42.9%	27.0%	hyper
chr3:33813273-33814104	75.22	73.4%	46.3%	27.0%	hyper
chr7:52964589-52965359	74.06	82.6%	55.6%	27.0%	hyper
chr4:3724809-3725422	66.56	75.1%	48.0%	27.0%	hyper
chr11:114647957-114648194	71.64	56.9%	29.8%	27.0%	hyper
chr6:30522799-30523141	86.01	88.5%	61.5%	27.0%	hyper
chr2:71134095-71134836	107.81	83.5%	56.5%	27.0%	hyper
chr14:56223568-56224041	64.75	75.5%	48.5%	27.0%	hyper
chr8:73309871-73310263	77.19	80.4%	53.4%	27.0%	hyper
chr14:32475104-32475476	123.68	57.5%	30.6%	27.0%	hyper
chr12:87418965-87419462	84.89	71.9%	44.9%	27.0%	hyper
chr7:51857056-51858580	126.45	52.7%	25.7%	27.0%	hyper
chr3:86803170-86803703	93.24	40.1%	13.1%	27.0%	hyper
chr6:65165344-65165747	86.22	80.5%	53.6%	27.0%	hyper
chr13:55516067-55516245	64.41	56.3%	29.3%	27.0%	hyper
chr11:53365100-53365677	123.06	83.6%	56.7%	26.9%	hyper
chr12:100878300-100879333	133.05	51.6%	24.7%	26.9%	hyper
chr11:4199526-4199942	58.17	76.1%	49.2%	26.9%	hyper
chr8:73334834-73335143	64.50	76.8%	49.9%	26.9%	hyper
chr9:66362096-66362549	64.50	75.7%	48.8%	26.9%	hyper
chr3:9544824-9545687	96.62	84.6%	57.7%	26.9%	hyper

chr4:151357238-151357763	76.61	71.9%	45.0%	26.9%	hyper
chr2:180012455-180013174	85.54	83.6%	56.8%	26.8%	hyper
chr11:97898955-97899568	96.90	83.2%	56.4%	26.8%	hyper
chr7:53917051-53917600	129.61	83.7%	56.9%	26.8%	hyper
chr1:88401554-88401922	64.96	41.2%	14.4%	26.8%	hyper
chr18:13190385-13190850	67.84	61.5%	34.7%	26.8%	hyper
chr19:42734006-42734476	153.50	75.7%	48.9%	26.8%	hyper
chr11:60198554-60198867	73.34	49.9%	23.1%	26.8%	hyper
chr7:105571609-105572323	72.65	76.4%	49.6%	26.8%	hyper
chr2:153448100-153448736	104.69	66.0%	39.3%	26.8%	hyper
chr11:115468146-115468403	88.93	87.4%	60.7%	26.8%	hyper
chr4:41044457-41044778	63.36	73.1%	46.3%	26.7%	hyper
chr5:130508348-130508692	70.14	86.5%	59.8%	26.7%	hyper
chr4:45335568-45335834	65.81	86.2%	59.5%	26.7%	hyper
chr13:55950222-55951134	105.07	83.3%	56.6%	26.7%	hyper
chr14:55503212-55503426	73.10	50.3%	23.6%	26.7%	hyper
chr9:42750187-42750934	127.22	65.0%	38.3%	26.7%	hyper
chr18:64173299-64174269	138.21	52.2%	25.5%	26.7%	hyper
chr5:114150244-114150866	88.12	80.5%	53.9%	26.7%	hyper
chr11:94295032-94295559	144.02	54.8%	28.1%	26.7%	hyper
chr1:134483359-134484141	89.68	76.7%	50.0%	26.7%	hyper
chr2:168645700-168646838	136.32	72.7%	46.0%	26.7%	hyper
chr7:50898544-50898991	93.95	83.4%	56.8%	26.7%	hyper
chr10:93733993-93734770	71.83	76.9%	50.3%	26.6%	hyper
chr6:35821166-35822056	111.68	77.5%	50.8%	26.6%	hyper
chr5:139407952-139408794	132.23	60.6%	34.0%	26.6%	hyper
chr2:167167083-167167987	77.82	88.6%	62.0%	26.6%	hyper
chr2:170114118-170115463	145.35	76.7%	50.1%	26.6%	hyper
chr4:153968970-153969275	83.40	72.7%	46.1%	26.6%	hyper
chr1:146167850-146168497	148.50	81.8%	55.2%	26.6%	hyper
chr1:82816715-82817132	66.91	69.4%	42.8%	26.6%	hyper
chr9:34679219-34679789	112.89	82.6%	56.0%	26.6%	hyper
chr5:125696058-125696776	183.33	51.2%	24.6%	26.6%	hyper
chr17:45942041-45942526	143.95	44.4%	17.8%	26.6%	hyper
chr16:48021723-48022248	80.11	79.5%	52.9%	26.6%	hyper
chr17:78631549-78632064	96.71	84.6%	58.1%	26.5%	hyper
chr18:36093231-36093860	87.55	63.3%	36.7%	26.5%	hyper
chr12:119758305-119759413	86.49	57.2%	30.7%	26.5%	hyper
chr4:150311839-150312368	103.95	82.5%	56.0%	26.5%	hyper
chr2:144219313-144219829	93.07	86.0%	59.5%	26.5%	hyper
chr15:84961845-84962260	130.40	61.3%	34.8%	26.5%	hyper
chr2:59681288-59681610	96.99	89.0%	62.5%	26.5%	hyper
chr18:25748789-25749307	151.84	56.0%	29.6%	26.5%	hyper
chr12:105956774-105957339	79.44	89.3%	62.8%	26.5%	hyper

chr10:62884349-62884844	138.69	88.4%	62.0%	26.5%	hyper
chr8:123407292-123407972	132.84	62.7%	36.2%	26.5%	hyper
chr16:30692101-30692962	165.75	80.2%	53.8%	26.4%	hyper
chr4:118716436-118717478	108.51	69.9%	43.5%	26.4%	hyper
chr8:10249084-10249470	150.93	70.9%	44.6%	26.4%	hyper
chr10:117351541-117352185	83.01	84.3%	57.9%	26.4%	hyper
chr8:124860571-124861615	444.71	48.8%	22.4%	26.4%	hyper
chr8:28231705-28232025	75.65	52.4%	26.0%	26.4%	hyper
chr11:118058984-118059675	82.69	82.9%	56.6%	26.3%	hyper
chr15:25375404-25375957	122.34	64.2%	37.9%	26.3%	hyper
chr19:47779581-47780615	74.20	84.0%	57.7%	26.3%	hyper
chr5:36083518-36083985	73.22	47.7%	21.4%	26.3%	hyper
chr9:122656248-122656723	95.27	80.4%	54.1%	26.3%	hyper
chr10:111304211-111305093	80.28	56.6%	30.3%	26.3%	hyper
chr17:34518802-34519358	61.01	82.6%	56.3%	26.3%	hyper
chr2:101819245-101819839	105.28	47.7%	21.5%	26.3%	hyper
chr1:75701215-75701860	102.72	66.3%	40.1%	26.3%	hyper
chr10:40151916-40152774	67.87	92.3%	66.0%	26.2%	hyper
chr19:57216314-57217562	96.32	83.5%	57.3%	26.2%	hyper
chr6:85373806-85374206	83.84	65.4%	39.2%	26.2%	hyper
chr13:47320013-47320660	83.63	49.5%	23.3%	26.2%	hyper
chr15:82841272-82841952	107.02	65.0%	38.9%	26.2%	hyper
chr3:105854909-105855384	101.43	76.3%	50.2%	26.2%	hyper
chr14:32485016-32485775	95.66	75.7%	49.6%	26.2%	hyper
chr16:30129449-30130025	91.63	68.1%	42.0%	26.1%	hyper
chr2:171833111-171833641	99.94	89.0%	62.9%	26.1%	hyper
chr3:101814502-101815228	79.83	64.9%	38.8%	26.1%	hyper
chr3:128941497-128941984	84.74	77.1%	51.0%	26.1%	hyper
chr9:72003285-72004022	57.70	73.8%	47.7%	26.1%	hyper
chr17:30068209-30068571	69.63	83.2%	57.1%	26.1%	hyper
chr19:59007841-59008889	92.62	63.1%	37.0%	26.1%	hyper
chr9:72701696-72702187	78.09	84.3%	58.2%	26.1%	hyper
chr10:79197762-79198198	324.77	75.5%	49.5%	26.1%	hyper
chr1:59219905-59220421	71.52	64.3%	38.3%	26.1%	hyper
chr6:48387994-48388871	422.41	45.1%	19.1%	26.0%	hyper
chr7:137894203-137894602	88.47	76.5%	50.5%	26.0%	hyper
chr3:135147635-135148152	109.27	75.2%	49.2%	26.0%	hyper
chr7:148274750-148274879	72.85	78.4%	52.4%	26.0%	hyper
chr5:75122364-75122946	69.93	76.5%	50.5%	26.0%	hyper
chr5:136209726-136210107	116.28	85.2%	59.2%	26.0%	hyper
chr4:148295401-148296206	101.25	82.2%	56.3%	26.0%	hyper
chr7:28277102-28277314	76.52	76.3%	50.4%	25.9%	hyper
chr9:63382057-63383206	81.11	67.1%	41.2%	25.9%	hyper
chr17:5849199-5849791	94.76	65.2%	39.3%	25.9%	hyper

chr15:8790136-8790425	88.62	91.0%	65.1%	25.9%	hyper
chr12:111792948-111793412	95.10	93.0%	67.1%	25.9%	hyper
chr3:79683120-79683489	73.31	84.2%	58.3%	25.9%	hyper
chr6:116528633-116528925	74.34	88.2%	62.4%	25.9%	hyper
chr8:68066541-68067285	178.54	84.0%	58.1%	25.9%	hyper
chr2:167002280-167003039	83.86	78.2%	52.3%	25.9%	hyper
chr15:102889861-102890264	89.35	47.6%	21.7%	25.9%	hyper
chr19:6041517-6041969	79.44	84.8%	59.0%	25.9%	hyper
chr18:10052318-10052935	79.27	64.4%	38.5%	25.9%	hyper
chr8:44688662-44689212	120.60	77.6%	51.7%	25.9%	hyper
chr7:28066903-28067263	60.50	76.2%	50.4%	25.8%	hyper
chr2:78122006-78122785	90.25	83.0%	57.1%	25.8%	hyper
chr5:69923232-69923595	66.74	80.3%	54.5%	25.8%	hyper
chr7:148192086-148193168	71.79	88.7%	62.9%	25.8%	hyper
chr17:93600126-93600477	67.20	40.3%	14.5%	25.8%	hyper
chr13:114377391-114378014	102.98	67.2%	41.5%	25.7%	hyper
chr2:44412774-44413769	83.12	63.3%	37.5%	25.7%	hyper
chr5:106038942-106039537	74.97	86.2%	60.5%	25.7%	hyper
chr11:44654709-44656167	75.49	73.5%	47.8%	25.7%	hyper
chr11:70028833-70029223	138.66	53.3%	27.6%	25.7%	hyper
chr18:30428994-30429263	93.73	73.4%	47.8%	25.7%	hyper
chr4:153764294-153764862	92.65	52.7%	27.1%	25.6%	hyper
chr5:65419471-65419822	79.56	85.2%	59.6%	25.6%	hyper
chr5:117931378-117931930	76.61	84.5%	58.8%	25.6%	hyper
chr2:166141605-166142442	124.06	58.2%	32.6%	25.6%	hyper
chr8:71661914-71662550	123.78	85.3%	59.7%	25.6%	hyper
chr12:58609755-58610353	90.03	81.3%	55.7%	25.6%	hyper
chr2:75828899-75829287	64.37	66.0%	40.4%	25.6%	hyper
chr2:11357855-11358212	95.40	86.9%	61.3%	25.6%	hyper
chr6:125549217-125549462	73.75	92.2%	66.6%	25.6%	hyper
chr9:103345760-103346400	108.09	67.7%	42.1%	25.5%	hyper
chr4:127096728-127097149	82.12	80.0%	54.5%	25.5%	hyper
chr7:52887987-52888323	148.09	68.3%	42.7%	25.5%	hyper
chr11:102852985-102853390	80.96	41.3%	15.8%	25.5%	hyper
chr4:150462956-150463253	72.30	83.8%	58.3%	25.5%	hyper
chr1:190806696-190806971	65.79	91.3%	65.8%	25.5%	hyper
chr4:152373616-152374294	131.83	82.3%	56.8%	25.5%	hyper
chr11:114036955-114038045	110.50	83.7%	58.2%	25.5%	hyper
chr18:86642977-86643580	77.69	78.1%	52.6%	25.5%	hyper
chr1:145617533-145617778	66.73	77.3%	51.8%	25.5%	hyper
chr4:83868677-83869492	71.96	88.0%	62.6%	25.5%	hyper
chr13:31707063-31707358	116.25	50.2%	24.7%	25.4%	hyper
chr2:22086168-22086830	115.78	80.1%	54.6%	25.4%	hyper
chr6:49223570-49224587	75.72	79.6%	54.2%	25.4%	hyper

chr10:81091261-81091606	70.80	80.7%	55.3%	25.4%	hyper
chr8:26059132-26059550	85.44	80.5%	55.1%	25.4%	hyper
chr13:3477491-3477905	159.67	55.9%	30.5%	25.4%	hyper
chr12:111186293-111187835	159.35	61.3%	35.9%	25.4%	hyper
chr2:143910986-143911396	71.81	83.8%	58.4%	25.4%	hyper
chr17:12411691-12412113	97.92	86.6%	61.2%	25.4%	hyper
chr10:67808171-67808931	93.36	81.9%	56.5%	25.4%	hyper
chr14:122852090-122852827	126.99	69.2%	43.8%	25.4%	hyper
chr8:73928811-73929522	90.59	63.0%	37.7%	25.3%	hyper
chr1:193040163-193040809	78.85	71.7%	46.3%	25.3%	hyper
chr9:63775068-63775725	282.91	50.0%	24.7%	25.3%	hyper
chr18:64178234-64179244	87.25	77.2%	51.9%	25.3%	hyper
chr10:95179792-95180259	86.26	91.3%	66.0%	25.3%	hyper
chr2:132333170-132334036	94.16	80.3%	55.0%	25.3%	hyper
chr8:28232239-28232378	86.59	57.5%	32.2%	25.3%	hyper
chr19:43645949-43646291	98.47	62.9%	37.6%	25.3%	hyper
chr10:41710457-41711981	99.07	58.6%	33.3%	25.2%	hyper
chr16:28927071-28927908	90.58	63.8%	38.6%	25.2%	hyper
chr17:10098969-10099649	144.62	41.1%	15.9%	25.2%	hyper
chr12:118266635-118267509	82.68	82.7%	57.4%	25.2%	hyper
chr6:24298057-24298798	82.46	69.8%	44.6%	25.2%	hyper
chr7:51806886-51807709	211.34	65.4%	40.3%	25.2%	hyper
chr14:104177521-104177819	67.76	65.2%	40.0%	25.1%	hyper
chr9:75230655-75231000	84.07	80.2%	55.0%	25.1%	hyper
chr17:81286827-81287263	63.61	83.3%	58.2%	25.1%	hyper
chr17:27250725-27251157	114.45	88.3%	63.2%	25.1%	hyper
chr13:15712273-15713182	131.55	89.6%	64.5%	25.1%	hyper
chr3:65197318-65197517	63.84	85.2%	60.1%	25.1%	hyper
chr2:6295297-6295664	74.43	78.0%	52.9%	25.1%	hyper
chr10:111095698-111096262	73.07	85.9%	60.8%	25.1%	hyper
chr3:142350957-142351342	135.81	81.5%	56.4%	25.1%	hyper
chr7:142657014-142657380	87.79	63.6%	38.5%	25.1%	hyper
chr5:138218953-138219448	71.77	82.3%	57.2%	25.1%	hyper
chr10:43608774-43609078	92.32	85.6%	60.6%	25.1%	hyper
chr14:55330531-55330769	90.31	87.5%	62.5%	25.0%	hyper
chr10:76713115-76714336	138.07	83.3%	58.3%	25.0%	hyper
chr14:8560503-8561146	111.59	79.4%	54.4%	25.0%	hyper
chr9:107516451-107517018	107.71	86.9%	61.9%	25.0%	hyper
chr7:26430147-26430915	66.37	75.7%	50.7%	25.0%	hyper
chr11:5394993-5395635	94.37	77.2%	52.2%	25.0%	hyper
chr9:62526177-62527084	183.93	63.6%	38.6%	25.0%	hyper
chr1:140468006-140468791	169.54	63.0%	38.0%	25.0%	hyper
chr18:38749564-38750007	93.52	84.1%	59.1%	25.0%	hyper
chr10:80642606-80643127	104.95	86.5%	61.5%	25.0%	hyper

chr16:90849699-90850223	72.78	78.4%	53.5%	25.0%	hyper
chr10:87323389-87323840	74.11	57.1%	32.1%	24.9%	hyper
chr5:61460063-61460178	70.18	96.1%	71.2%	24.9%	hyper
chr15:97966088-97966358	59.77	88.2%	63.3%	24.9%	hyper
chr4:124344632-124345709	140.75	61.9%	37.0%	24.9%	hyper
chr4:153978242-153978684	83.69	78.4%	53.5%	24.9%	hyper
chr2:147849058-147849438	63.19	73.3%	48.4%	24.9%	hyper
chr5:65758918-65759438	80.35	87.2%	62.3%	24.9%	hyper
chr13:108453432-108454568	97.22	57.0%	32.2%	24.8%	hyper
chr3:108684075-108684565	95.63	79.3%	54.4%	24.8%	hyper
chr1:89371617-89372259	215.45	36.0%	11.2%	24.8%	hyper
chr9:63602401-63603316	168.22	74.6%	49.7%	24.8%	hyper
chr5:24481953-24482907	88.62	74.2%	49.4%	24.8%	hyper
chr14:61623709-61624298	71.71	38.0%	13.2%	24.8%	hyper
chr4:139195967-139196611	98.95	80.7%	56.0%	24.8%	hyper
chr5:129099014-129099686	94.15	78.1%	53.3%	24.8%	hyper
chr7:150481280-150481610	155.14	50.8%	26.1%	24.7%	hyper
chr1:180501898-180502664	97.24	79.3%	54.6%	24.7%	hyper
chr16:20701731-20702938	96.31	49.9%	25.2%	24.7%	hyper
chr11:61263032-61263505	80.65	84.8%	60.1%	24.7%	hyper
chr8:63704673-63705281	110.78	81.2%	56.5%	24.7%	hyper
chr3:69529390-69530253	65.32	90.6%	65.9%	24.7%	hyper
chr2:73160999-73161557	90.48	82.4%	57.7%	24.7%	hyper
chr5:67529680-67530504	78.52	60.3%	35.6%	24.7%	hyper
chr8:126416615-126416889	79.38	53.8%	29.1%	24.7%	hyper
chr19:3904259-3904600	76.58	79.0%	54.3%	24.7%	hyper
chr5:117627706-117628918	116.83	86.2%	61.6%	24.7%	hyper
chr2:17821181-17821526	70.03	85.6%	60.9%	24.7%	hyper
chr10:10871443-10871705	66.93	89.7%	65.0%	24.7%	hyper
chr5:129837005-129837265	64.94	59.5%	34.8%	24.7%	hyper
chr13:24613355-24613926	83.51	87.3%	62.7%	24.7%	hyper
chr7:26002593-26002783	59.38	83.8%	59.2%	24.6%	hyper
chr11:116899342-116900292	80.92	65.1%	40.5%	24.6%	hyper
chr17:6940931-6941564	96.63	74.8%	50.2%	24.6%	hyper
chr11:119045151-119045702	79.56	82.5%	57.9%	24.6%	hyper
chr13:38502240-38502931	82.31	74.0%	49.4%	24.6%	hyper
chr11:94104505-94104840	116.06	55.8%	31.2%	24.6%	hyper
chr10:60257254-60257825	91.35	56.2%	31.7%	24.6%	hyper
chr9:120617893-120618461	81.30	67.6%	43.0%	24.6%	hyper
chr4:149507100-149507753	59.97	62.5%	38.0%	24.6%	hyper
chr3:153788068-153788365	75.39	84.8%	60.2%	24.5%	hyper
chr4:149467829-149468913	83.71	74.3%	49.7%	24.5%	hyper
chr6:52053517-52054806	95.06	86.2%	61.6%	24.5%	hyper
chr13:72954239-72954833	72.96	59.9%	35.3%	24.5%	hyper

chr1:157338344-157339336	86.19	75.7%	51.2%	24.5%	hyper
chr18:61803235-61803777	80.29	54.5%	30.0%	24.5%	hyper
chr14:56380507-56380930	128.73	40.6%	16.1%	24.5%	hyper
chr8:67236834-67237160	125.73	89.0%	64.5%	24.5%	hyper
chr10:57514366-57514805	137.70	43.2%	18.7%	24.4%	hyper
chr11:63702619-63702772	95.40	83.6%	59.2%	24.4%	hyper
chr3:138368976-138369686	79.81	76.6%	52.2%	24.4%	hyper
chr13:34340184-34341500	88.57	62.9%	38.5%	24.4%	hyper
chr5:124022234-124022691	91.61	63.4%	39.0%	24.4%	hyper
chr4:93791098-93791531	57.58	79.3%	55.0%	24.4%	hyper
chr2:6022784-6023798	91.17	88.4%	64.0%	24.4%	hyper
chr2:172759340-172759801	155.47	58.1%	33.8%	24.3%	hyper
chr4:116502237-116503120	82.03	62.7%	38.4%	24.3%	hyper
chr2:172682801-172683281	106.30	85.7%	61.4%	24.3%	hyper
chr1:138678412-138679136	112.42	75.4%	51.1%	24.3%	hyper
chr4:150132691-150133081	104.19	53.0%	28.7%	24.3%	hyper
chr5:143653190-143653867	69.05	80.7%	56.4%	24.3%	hyper
chr7:147973535-147973822	84.61	80.3%	56.0%	24.3%	hyper
chr10:110196760-110197167	137.00	87.3%	63.0%	24.3%	hyper
chr11:94426344-94426681	93.80	51.5%	27.2%	24.3%	hyper
chr18:65092959-65093669	113.77	57.3%	33.0%	24.3%	hyper
chr16:5007989-5008567	69.40	85.1%	60.8%	24.3%	hyper
chr18:64084059-64084745	81.05	70.1%	45.8%	24.3%	hyper
chr11:30917179-30917613	65.95	82.6%	58.3%	24.3%	hyper
chr4:125092602-125093018	96.87	75.9%	51.6%	24.3%	hyper
chr4:133568151-133569516	101.01	55.9%	31.6%	24.3%	hyper
chr5:115594001-115595018	100.03	84.8%	60.6%	24.3%	hyper
chr3:89029673-89030383	88.41	35.6%	11.3%	24.2%	hyper
chr2:158625885-158627235	115.47	86.8%	62.6%	24.2%	hyper
chr10:128158210-128158582	67.82	83.9%	59.7%	24.2%	hyper
chr3:84548485-84548742	66.08	80.5%	56.4%	24.2%	hyper
chr1:80236564-80237093	78.82	76.4%	52.3%	24.2%	hyper
chr7:29997949-29998493	72.05	63.5%	39.3%	24.2%	hyper
chr6:82863404-82863753	81.37	85.4%	61.2%	24.2%	hyper
chr11:84006594-84007596	90.07	87.2%	63.0%	24.2%	hyper
chr7:26489381-26490333	188.57	38.0%	13.9%	24.2%	hyper
chr18:75082914-75083690	94.35	87.2%	63.0%	24.1%	hyper
chr5:148079766-148080598	165.48	44.2%	20.1%	24.1%	hyper
chr10:129849444-129849658	129.54	86.6%	62.4%	24.1%	hyper
chr5:140740108-140740522	75.24	76.6%	52.5%	24.1%	hyper
chr1:21086994-21087549	114.46	84.2%	60.1%	24.1%	hyper
chr6:125313074-125313725	91.09	64.9%	40.8%	24.1%	hyper
chr8:91496929-91497489	82.18	77.6%	53.5%	24.1%	hyper
chr6:144212752-144213405	86.27	77.9%	53.8%	24.1%	hyper

chr12:73564853-73565318	76.55	85.4%	61.3%	24.1%	hyper
chr11:48639471-48639827	111.52	46.8%	22.7%	24.1%	hyper
chr2:74564000-74564316	77.84	52.3%	28.2%	24.1%	hyper
chr18:61104671-61105055	97.36	76.2%	52.2%	24.1%	hyper
chr7:31658617-31658960	61.80	79.7%	55.6%	24.1%	hyper
chr1:167565599-167566481	87.42	72.5%	48.5%	24.1%	hyper
chr10:94773904-94774389	73.26	56.5%	32.5%	24.1%	hyper
chr1:193809076-193809500	138.98	85.7%	61.7%	24.1%	hyper
chr10:80997859-80998416	108.01	62.9%	38.9%	24.1%	hyper
chr12:85105018-85105486	104.26	69.1%	45.1%	24.0%	hyper
chr2:105071605-105072267	61.23	75.8%	51.8%	24.0%	hyper
chr11:105850073-105850305	118.53	81.7%	57.7%	24.0%	hyper
chr16:32238407-32238921	94.00	88.8%	64.7%	24.0%	hyper
chr16:29837319-29837762	110.67	66.9%	42.9%	24.0%	hyper
chr8:88378216-88378602	74.61	81.9%	57.9%	24.0%	hyper
chr2:33772099-33772439	67.70	87.4%	63.4%	24.0%	hyper
chr4:129065980-129066304	65.10	87.1%	63.1%	24.0%	hyper
chr14:56434090-56434667	128.02	73.2%	49.2%	24.0%	hyper
chr13:48878002-48878606	157.73	81.3%	57.4%	24.0%	hyper
chr7:31639938-31640405	67.71	83.7%	59.8%	23.9%	hyper
chr9:29769692-29770666	143.95	55.4%	31.5%	23.9%	hyper
chr11:68948512-68948939	97.70	84.8%	60.9%	23.9%	hyper
chr5:92522253-92522998	75.50	77.9%	54.0%	23.9%	hyper
chr17:84965826-84966301	88.70	86.0%	62.1%	23.9%	hyper
chr10:28916568-28917185	229.11	87.7%	63.8%	23.9%	hyper
chr9:97091307-97091832	94.48	89.4%	65.5%	23.9%	hyper
chr9:35624933-35625294	62.89	83.1%	59.2%	23.9%	hyper
chr11:107564756-107565372	99.99	80.9%	57.0%	23.9%	hyper
chr13:38033390-38033995	85.98	83.7%	59.8%	23.9%	hyper
chr7:4434397-4434994	68.20	86.5%	62.6%	23.9%	hyper
chr1:183783324-183783695	61.99	56.7%	32.8%	23.9%	hyper
chr12:3941833-3942325	91.39	70.6%	46.8%	23.9%	hyper
chr5:111849752-111850042	91.30	58.3%	34.4%	23.9%	hyper
chr2:4861933-4862662	101.54	85.5%	61.7%	23.9%	hyper
chr13:91880797-91881264	132.92	41.6%	17.8%	23.9%	hyper
chr12:88030038-88030321	84.61	88.0%	64.2%	23.8%	hyper
chr13:96764716-96765205	101.85	74.7%	50.9%	23.8%	hyper
chr13:98344119-98344970	99.20	81.9%	58.1%	23.8%	hyper
chr8:11437736-11438504	126.39	74.3%	50.5%	23.8%	hyper
chr19:55769166-55769739	69.39	43.9%	20.0%	23.8%	hyper
chr11:100274930-100275365	73.65	72.7%	48.9%	23.8%	hyper
chr13:59175359-59175549	63.05	68.0%	44.1%	23.8%	hyper
chr10:44729376-44729627	74.98	71.5%	47.7%	23.8%	hyper
chr18:76772427-76772881	82.43	65.3%	41.5%	23.8%	hyper

chr11:71842650-71843015	94.79	73.7%	49.9%	23.8%	hyper
chr13:38022600-38023145	178.86	87.1%	63.3%	23.8%	hyper
chr3:127453146-127453657	72.51	73.6%	49.8%	23.8%	hyper
chr10:80870954-80871435	74.95	68.9%	45.2%	23.7%	hyper
chr10:59383863-59384262	97.38	51.2%	27.5%	23.7%	hyper
chr1:182214270-182214464	76.15	54.2%	30.5%	23.7%	hyper
chr10:6382293-6382828	143.64	83.7%	60.0%	23.7%	hyper
chr8:64010893-64011317	68.81	80.8%	57.1%	23.7%	hyper
chr17:66199213-66199496	106.20	88.3%	64.7%	23.7%	hyper
chr16:42955789-42955991	118.39	89.2%	65.6%	23.7%	hyper
chr10:62488631-62489358	66.91	78.2%	54.5%	23.7%	hyper
chr3:56175569-56175925	71.97	83.8%	60.2%	23.6%	hyper
chr5:97516363-97516798	103.60	84.5%	60.8%	23.6%	hyper
chr7:144063448-144063895	100.20	64.6%	40.9%	23.6%	hyper
chr9:72704445-72704903	115.19	84.9%	61.3%	23.6%	hyper
chr5:124799274-124799744	128.32	85.5%	61.9%	23.6%	hyper
chr12:106972538-106973124	89.31	81.8%	58.2%	23.6%	hyper
chr5:139992139-139992625	91.17	82.4%	58.8%	23.6%	hyper
chr9:24132783-24133056	98.34	86.3%	62.7%	23.6%	hyper
chr8:90307972-90308453	104.05	63.4%	39.8%	23.6%	hyper
chr13:51848252-51848878	65.12	75.4%	51.8%	23.6%	hyper
chr17:32975893-32976378	64.98	73.0%	49.4%	23.6%	hyper
chr2:115404974-115405886	74.69	56.9%	33.3%	23.6%	hyper
chr10:97375927-97376644	118.49	87.8%	64.2%	23.6%	hyper
chr4:84928893-84929562	94.45	85.9%	62.3%	23.6%	hyper
chr3:95363349-95363851	162.97	86.6%	63.0%	23.6%	hyper
chr9:44370960-44371601	68.29	72.5%	48.9%	23.6%	hyper
chr3:95636036-95636600	95.16	91.3%	67.7%	23.5%	hyper
chr4:139942136-139943259	129.85	86.0%	62.4%	23.5%	hyper
chr5:37373158-37373630	109.39	82.1%	58.6%	23.5%	hyper
chr12:71328020-71328962	197.34	40.4%	16.9%	23.5%	hyper
chr13:39117939-39118651	72.16	94.1%	70.6%	23.5%	hyper
chr2:172739426-172739710	59.42	78.6%	55.2%	23.5%	hyper
chr11:49078159-49078495	58.24	85.9%	62.4%	23.5%	hyper
chr16:91232100-91233291	232.15	41.8%	18.4%	23.5%	hyper
chr4:57425340-57426500	87.95	65.8%	42.3%	23.5%	hyper
chr3:86354083-86355266	151.50	45.9%	22.5%	23.5%	hyper
chr18:3499113-3499302	65.18	76.7%	53.3%	23.4%	hyper
chr6:145494660-145495165	70.00	51.4%	27.9%	23.4%	hyper
chr8:10793933-10794799	76.78	87.2%	63.7%	23.4%	hyper
chr17:64182586-64183094	81.83	87.5%	64.1%	23.4%	hyper
chr6:113308644-113309460	87.19	92.8%	69.4%	23.4%	hyper
chr19:24258344-24258876	75.04	86.1%	62.7%	23.4%	hyper
chr11:74115434-74115990	122.47	83.7%	60.3%	23.4%	hyper

chr3:34793190-34793820	75.33	88.2%	64.8%	23.4%	hyper
chr15:78719815-78720340	60.50	64.6%	41.2%	23.4%	hyper
chr13:43786356-43786930	60.82	69.6%	46.2%	23.4%	hyper
chr12:56961766-56962204	80.43	89.5%	66.2%	23.4%	hyper
chr9:77811925-77812241	72.31	82.3%	58.9%	23.4%	hyper
chr2:25280567-25280949	86.98	81.9%	58.6%	23.4%	hyper
chr18:61228464-61228781	71.32	76.0%	52.7%	23.4%	hyper
chr11:17108414-17108700	70.55	92.1%	68.8%	23.4%	hyper
chr5:112264855-112266022	115.62	62.6%	39.2%	23.4%	hyper
chr4:28227063-28227620	97.49	74.9%	51.5%	23.3%	hyper
chr6:54839870-54840604	82.88	78.6%	55.3%	23.3%	hyper
chr15:80416857-80417546	107.26	78.6%	55.3%	23.3%	hyper
chr11:96338361-96338945	103.93	90.2%	66.9%	23.3%	hyper
chr17:86080044-86080530	107.10	54.1%	30.7%	23.3%	hyper
chr2:165174873-165175418	82.37	67.8%	44.5%	23.3%	hyper
chr11:44119523-44119976	85.03	83.9%	60.6%	23.3%	hyper
chr19:55701071-55701907	117.81	89.2%	65.9%	23.3%	hyper
chr3:126997110-126997376	70.26	87.7%	64.4%	23.3%	hyper
chr1:130336729-130337421	81.79	69.9%	46.6%	23.3%	hyper
chr18:36483386-36484009	107.02	68.3%	45.0%	23.3%	hyper
chr2:167385052-167385501	61.61	59.8%	36.5%	23.3%	hyper
chr7:106554438-106554929	76.01	66.1%	42.8%	23.3%	hyper
chr18:76305414-76305883	162.53	83.2%	59.9%	23.3%	hyper
chr11:108796718-108797685	77.62	76.9%	53.7%	23.2%	hyper
chr4:9420973-9421163	66.24	80.4%	57.2%	23.2%	hyper
chr8:121269078-121269783	76.81	90.6%	67.4%	23.2%	hyper
chr10:122320820-122321602	59.53	61.1%	37.9%	23.2%	hyper
chr2:179319646-179320177	79.99	84.8%	61.6%	23.2%	hyper
chr1:159388897-159389356	129.13	40.0%	16.8%	23.2%	hyper
chr2:33431053-33431191	57.22	65.1%	41.9%	23.2%	hyper
chr17:28995304-28995687	93.15	78.8%	55.6%	23.2%	hyper
chr10:42874479-42875539	120.58	70.1%	46.9%	23.2%	hyper
chr9:60674035-60674390	58.73	89.1%	65.9%	23.2%	hyper
chr3:51025786-51026567	74.56	82.4%	59.2%	23.2%	hyper
chr9:14050690-14051104	60.77	61.9%	38.8%	23.2%	hyper
chr2:169919826-169920452	111.20	82.6%	59.5%	23.1%	hyper
chr1:72284301-72284987	66.95	76.1%	53.0%	23.1%	hyper
chr16:92475463-92476098	119.58	55.5%	32.4%	23.1%	hyper
chr7:109437703-109438326	70.76	81.2%	58.1%	23.1%	hyper
chr1:89023684-89024330	241.75	69.0%	45.9%	23.1%	hyper
chr1:108609344-108609850	77.84	67.0%	43.9%	23.1%	hyper
chr3:94976010-94976718	263.03	44.5%	21.4%	23.1%	hyper
chr16:90712961-90713392	68.88	73.5%	50.4%	23.1%	hyper
chr15:76346668-76347120	83.32	83.9%	60.8%	23.1%	hyper

chr5:52759595-52759927	94.04	68.5%	45.4%	23.0%	hyper
chr13:23531767-23532126	69.70	63.7%	40.7%	23.0%	hyper
chr15:88481461-88481849	81.86	57.0%	34.0%	23.0%	hyper
chr6:39296910-39297415	65.32	60.0%	37.0%	23.0%	hyper
chr11:99854817-99855240	77.06	75.2%	52.1%	23.0%	hyper
chr8:8751043-8751461	85.83	91.4%	68.4%	23.0%	hyper
chr1:17081732-17082304	186.00	89.8%	66.8%	23.0%	hyper
chr11:102639062-102639563	78.48	65.8%	42.8%	23.0%	hyper
chr3:146089317-146090045	88.05	85.9%	62.9%	23.0%	hyper
chr12:107122636-107123272	80.46	79.9%	56.9%	23.0%	hyper
chr13:114661022-114661863	87.26	84.9%	61.9%	23.0%	hyper
chr11:97445466-97446044	64.10	87.4%	64.4%	23.0%	hyper
chr10:120979633-120980040	94.45	83.6%	60.7%	23.0%	hyper
chr7:35195922-35196505	87.64	87.4%	64.4%	23.0%	hyper
chr8:119727721-119728712	228.19	87.0%	64.1%	23.0%	hyper
chr13:42288445-42288863	64.82	83.3%	60.3%	23.0%	hyper
chr5:86434756-86434999	88.93	88.0%	65.1%	23.0%	hyper
chr10:41917209-41917728	207.27	83.0%	60.0%	23.0%	hyper
chr4:114751244-114752324	72.31	65.2%	42.3%	23.0%	hyper
chr1:34093309-34093859	77.35	84.1%	61.2%	23.0%	hyper
chr5:117984058-117984488	77.16	48.1%	25.1%	23.0%	hyper
chr8:63726230-63726897	76.22	82.8%	59.8%	22.9%	hyper
chr5:106990187-106990322	71.23	60.2%	37.3%	22.9%	hyper
chr1:187308178-187308821	115.49	88.1%	65.2%	22.9%	hyper
chr2:154471294-154471608	110.93	81.6%	58.6%	22.9%	hyper
chr10:29678916-29679417	146.96	87.6%	64.7%	22.9%	hyper
chr14:76816192-76816514	86.42	69.1%	46.2%	22.9%	hyper
chr10:121336577-121337031	82.55	78.2%	55.3%	22.9%	hyper
chr12:86026705-86027689	68.00	84.0%	61.1%	22.9%	hyper
chr4:133969284-133969461	59.75	80.9%	58.0%	22.9%	hyper
chr2:163307816-163308276	107.60	36.7%	13.8%	22.9%	hyper
chr7:31250804-31251368	164.01	37.5%	14.6%	22.9%	hyper
chr9:96844894-96845318	96.03	81.7%	58.9%	22.8%	hyper
chr12:72196255-72197555	77.35	88.6%	65.7%	22.8%	hyper
chr15:82911394-82912030	74.44	77.9%	55.0%	22.8%	hyper
chr13:46853486-46854844	125.14	82.0%	59.1%	22.8%	hyper
chr9:57357758-57358309	68.11	81.2%	58.4%	22.8%	hyper
chr5:121233139-121233546	83.29	79.2%	56.4%	22.8%	hyper
chr12:81012117-81013102	104.55	71.8%	49.0%	22.8%	hyper
chr10:82482636-82482875	84.40	63.4%	40.7%	22.8%	hyper
chr14:120788061-120788917	171.18	33.6%	10.8%	22.8%	hyper
chr8:124993740-124994488	87.66	85.3%	62.6%	22.8%	hyper
chr6:47893375-47893595	63.96	56.6%	33.8%	22.8%	hyper
chr17:66305056-66306620	115.38	84.1%	61.3%	22.8%	hyper

chr18:61225022-61225835	78.32	49.3%	26.6%	22.7%	hyper
chr10:42649752-42650359	66.85	86.7%	63.9%	22.7%	hyper
chr1:91647503-91647844	89.37	75.5%	52.8%	22.7%	hyper
chr16:85145789-85146712	79.87	92.5%	69.8%	22.7%	hyper
chr18:80366297-80367149	144.92	62.6%	39.9%	22.7%	hyper
chr9:43855567-43856975	83.47	83.7%	61.0%	22.7%	hyper
chr8:113329436-113329852	68.90	75.9%	53.2%	22.7%	hyper
chr3:28476263-28476574	104.60	81.8%	59.1%	22.7%	hyper
chr7:6528466-6529042	95.84	85.5%	62.8%	22.7%	hyper
chr8:106771132-106771738	80.48	75.7%	53.0%	22.7%	hyper
chr2:119630442-119631561	93.51	43.2%	20.5%	22.7%	hyper
chr4:147373363-147374484	81.29	79.1%	56.4%	22.7%	hyper
chr10:12396403-12397067	72.22	89.9%	67.2%	22.6%	hyper
chr6:127650488-127651824	99.58	76.5%	53.8%	22.6%	hyper
chr4:94419178-94419801	66.94	79.1%	56.5%	22.6%	hyper
chr12:86542290-86542500	59.13	67.6%	45.0%	22.6%	hyper
chr19:45770138-45770706	73.54	76.7%	54.1%	22.6%	hyper
chr10:126246956-126247540	127.56	79.0%	56.4%	22.6%	hyper
chr3:63843225-63843541	95.22	85.5%	63.0%	22.6%	hyper
chr5:149520147-149520962	68.33	71.8%	49.3%	22.5%	hyper
chr6:72606706-72607315	60.65	83.7%	61.1%	22.5%	hyper
chr11:5697042-5697545	78.46	82.3%	59.8%	22.5%	hyper
chr5:104082460-104083278	87.32	72.5%	50.0%	22.5%	hyper
chr6:126484574-126484914	115.67	72.8%	50.3%	22.5%	hyper
chr15:63506719-63507483	80.80	89.0%	66.5%	22.5%	hyper
chr4:125923594-125924274	62.37	82.2%	59.7%	22.5%	hyper
chr19:37765321-37765560	140.41	34.0%	11.5%	22.5%	hyper
chr4:117825657-117826193	124.43	88.5%	66.0%	22.5%	hyper
chr6:85336726-85337105	63.84	68.2%	45.8%	22.4%	hyper
chr15:80967859-80968347	66.08	54.5%	32.1%	22.4%	hyper
chr7:147862195-147862418	79.42	35.9%	13.4%	22.4%	hyper
chr11:63769675-63771299	122.60	58.2%	35.8%	22.4%	hyper
chr13:38694526-38695149	67.82	87.0%	64.6%	22.4%	hyper
chr11:108810346-108811241	122.60	90.2%	67.8%	22.4%	hyper
chr8:74512528-74512971	70.77	45.7%	23.3%	22.4%	hyper
chr8:82353425-82353703	69.10	78.5%	56.1%	22.4%	hyper
chr17:27281633-27282001	73.28	88.9%	66.5%	22.4%	hyper
chr18:67019491-67019879	76.06	60.0%	37.6%	22.4%	hyper
chr8:127777838-127778436	98.93	67.7%	45.4%	22.4%	hyper
chr19:9021542-9022076	72.24	90.5%	68.1%	22.4%	hyper
chr8:59888760-59889406	117.24	50.6%	28.2%	22.4%	hyper
chr8:10899833-10900540	80.52	70.5%	48.1%	22.3%	hyper
chr12:5050749-5051275	78.76	82.9%	60.5%	22.3%	hyper
chr10:61914187-61914973	94.93	78.5%	56.2%	22.3%	hyper

chr16:92468411-92468768	59.51	80.5%	58.2%	22.3%	hyper
chr1:135331374-135331790	87.54	61.4%	39.2%	22.3%	hyper
chr11:28969443-28969774	83.02	77.5%	55.2%	22.3%	hyper
chr6:122513110-122513678	96.79	79.8%	57.5%	22.3%	hyper
chr9:96337251-96337716	73.04	60.7%	38.4%	22.3%	hyper
chr17:46657374-46657774	68.11	73.5%	51.3%	22.3%	hyper
chr5:147531605-147532266	116.14	86.5%	64.2%	22.2%	hyper
chr2:60390849-60391113	76.37	51.3%	29.0%	22.2%	hyper
chr15:76004856-76005224	86.13	90.2%	67.9%	22.2%	hyper
chr14:61325576-61325840	76.70	82.8%	60.6%	22.2%	hyper
chr8:116162286-116162581	56.65	85.3%	63.0%	22.2%	hyper
chr11:119872794-119873431	93.26	81.1%	58.8%	22.2%	hyper
chr19:46509452-46510022	79.47	77.2%	55.0%	22.2%	hyper
chr19:37507291-37507847	81.73	76.1%	53.9%	22.2%	hyper
chr3:66023593-66024140	93.45	41.8%	19.6%	22.2%	hyper
chr11:68246148-68246397	74.41	63.2%	41.1%	22.2%	hyper
chr15:85337985-85338351	62.79	81.0%	58.9%	22.2%	hyper
chr8:129775777-129776347	91.46	85.5%	63.4%	22.1%	hyper
chr16:35920309-35920878	68.84	83.0%	60.9%	22.1%	hyper
chr9:45743927-45744341	73.68	41.0%	18.9%	22.1%	hyper
chr10:42293263-42294335	70.39	51.8%	29.7%	22.1%	hyper
chr10:117100737-117101367	71.60	69.6%	47.5%	22.1%	hyper
chr17:43393338-43393805	92.45	86.5%	64.4%	22.1%	hyper
chr7:138741192-138741932	75.42	64.7%	42.7%	22.1%	hyper
chr14:61152407-61152863	108.72	79.1%	57.0%	22.1%	hyper
chr5:142985452-142985996	66.06	42.3%	20.2%	22.1%	hyper
chr19:48059826-48060423	72.43	79.9%	57.8%	22.1%	hyper
chr12:104553034-104553537	127.46	64.1%	42.0%	22.0%	hyper
chr19:23731825-23732552	89.14	76.2%	54.1%	22.0%	hyper
chr17:66289705-66289832	55.86	87.3%	65.2%	22.0%	hyper
chr1:69703523-69703976	81.90	70.8%	48.8%	22.0%	hyper
chr4:140147769-140147994	72.52	84.1%	62.0%	22.0%	hyper
chr5:72655316-72655932	99.86	73.2%	51.2%	22.0%	hyper
chr19:6121712-6122132	64.88	53.2%	31.1%	22.0%	hyper
chr5:108146908-108147700	59.62	74.0%	52.0%	22.0%	hyper
chr10:127036675-127037461	123.93	76.2%	54.2%	22.0%	hyper
chr14:60245599-60246377	64.55	63.6%	41.6%	22.0%	hyper
chr11:97556787-97557163	68.52	79.9%	57.9%	22.0%	hyper
chr17:29693507-29693978	98.42	84.3%	62.3%	22.0%	hyper
chr4:139560544-139561096	72.65	58.0%	36.0%	22.0%	hyper
chr3:9158996-9159640	78.56	52.4%	30.4%	22.0%	hyper
chr17:35154731-35154891	60.54	80.8%	58.8%	22.0%	hyper
chr10:44719902-44720497	98.83	70.7%	48.7%	22.0%	hyper
chr12:105934923-105935367	81.03	82.8%	60.8%	22.0%	hyper

chr10:43443124-43443663	103.77	39.1%	17.2%	21.9%	hyper
chr11:102217890-102218476	114.99	79.4%	57.5%	21.9%	hyper
chr4:57650902-57651446	117.19	78.4%	56.5%	21.9%	hyper
chr3:52070595-52070889	67.55	43.4%	21.5%	21.9%	hyper
chr4:155165951-155166417	89.90	88.2%	66.3%	21.9%	hyper
chr13:108662920-108663746	68.85	85.1%	63.2%	21.9%	hyper
chr9:120640980-120641601	71.18	53.1%	31.2%	21.9%	hyper
chr6:24464803-24465275	64.45	50.7%	28.8%	21.9%	hyper
chr17:79293671-79295137	105.26	75.0%	53.1%	21.9%	hyper
chr14:52608848-52609340	73.18	75.5%	53.6%	21.9%	hyper
chr12:52276394-52276733	75.41	82.5%	60.6%	21.9%	hyper
chr5:37113779-37115016	87.95	88.5%	66.6%	21.9%	hyper
chr10:126394056-126394363	69.11	38.4%	16.5%	21.9%	hyper
chr15:102986159-102986548	115.48	81.0%	59.2%	21.9%	hyper
chr5:13879511-13880234	116.88	81.6%	59.7%	21.8%	hyper
chr10:86193710-86194242	82.12	40.1%	18.2%	21.8%	hyper
chr9:56746871-56747490	82.50	90.9%	69.1%	21.8%	hyper
chr1:181127803-181129014	61.19	84.7%	62.9%	21.8%	hyper
chr11:35610726-35611148	95.58	73.5%	51.7%	21.8%	hyper
chr6:83664659-83664893	91.07	35.0%	13.3%	21.8%	hyper
chr1:134719445-134720265	98.17	87.4%	65.6%	21.8%	hyper
chr11:49403109-49403751	193.15	85.9%	64.2%	21.8%	hyper
chr10:127001091-127001544	62.29	70.6%	48.8%	21.8%	hyper
chr6:134743474-134744132	63.10	71.6%	49.9%	21.7%	hyper
chr2:80997423-80997784	80.91	86.6%	64.8%	21.7%	hyper
chr19:12321876-12322262	104.86	86.3%	64.5%	21.7%	hyper
chr11:84332644-84333069	89.85	50.6%	28.8%	21.7%	hyper
chr3:81949308-81949647	86.37	36.8%	15.0%	21.7%	hyper
chr7:105878281-105878743	109.25	59.1%	37.4%	21.7%	hyper
chr1:165861254-165861544	63.11	67.1%	45.4%	21.7%	hyper
chr1:36001284-36002222	75.32	79.5%	57.8%	21.7%	hyper
chr2:29050993-29051328	91.02	46.1%	24.4%	21.7%	hyper
chr5:139296867-139297159	85.42	64.0%	42.3%	21.7%	hyper
chr10:79976503-79977367	154.97	71.6%	49.9%	21.7%	hyper
chr5:116052592-116052896	65.26	87.6%	65.9%	21.7%	hyper
chr18:75582931-75583247	93.21	84.3%	62.6%	21.7%	hyper
chr15:89246899-89247470	82.93	85.8%	64.1%	21.7%	hyper
chr5:117690695-117691301	124.58	42.9%	21.2%	21.7%	hyper
chr9:108856210-108856568	76.45	62.3%	40.7%	21.7%	hyper
chr13:107958666-107959036	73.79	67.1%	45.5%	21.7%	hyper
chr10:61128047-61128375	58.04	92.1%	70.4%	21.7%	hyper
chr16:91156270-91157008	100.42	81.9%	60.3%	21.6%	hyper
chr15:35846613-35847040	89.60	81.7%	60.1%	21.6%	hyper
chr6:31546422-31547644	106.32	61.7%	40.1%	21.6%	hyper

chr18:82741129-82741340	65.06	82.0%	60.4%	21.6%	hyper
chr4:138134311-138134655	86.50	85.3%	63.7%	21.6%	hyper
chr9:119214019-119214528	68.42	87.4%	65.8%	21.6%	hyper
chr7:35537983-35538645	177.04	81.7%	60.0%	21.6%	hyper
chr5:123832415-123832981	68.44	83.3%	61.7%	21.6%	hyper
chr15:98626660-98627388	73.19	60.9%	39.3%	21.6%	hyper
chr1:138696965-138697319	79.39	91.0%	69.4%	21.6%	hyper
chr4:85129168-85129530	73.11	85.8%	64.2%	21.6%	hyper
chr14:59486254-59487060	71.95	86.1%	64.5%	21.6%	hyper
chr10:79382809-79383539	107.33	54.6%	33.0%	21.6%	hyper
chr15:99358058-99358530	97.77	67.4%	45.8%	21.6%	hyper
chr12:114054801-114055414	66.22	76.4%	54.8%	21.6%	hyper
chr10:66944845-66945381	69.92	85.2%	63.6%	21.5%	hyper
chr3:88990959-88991286	85.98	81.5%	59.9%	21.5%	hyper
chr11:17360956-17361270	84.95	55.2%	33.7%	21.5%	hyper
chr11:74671143-74671414	71.32	87.3%	65.8%	21.5%	hyper
chr10:22531874-22532521	79.46	79.5%	58.0%	21.5%	hyper
chr7:139158709-139159377	332.90	53.6%	32.2%	21.5%	hyper
chr3:18871881-18872531	90.09	83.6%	62.1%	21.5%	hyper
chr9:61421304-61421799	66.55	74.1%	52.6%	21.5%	hyper
chr14:70408834-70409366	82.90	73.0%	51.5%	21.5%	hyper
chr15:82038214-82038624	73.33	49.1%	27.7%	21.5%	hyper
chr1:40381923-40383583	313.07	43.0%	21.6%	21.5%	hyper
chr10:59570816-59571512	83.86	58.2%	36.8%	21.4%	hyper
chr3:97494897-97495488	67.01	81.5%	60.0%	21.4%	hyper
chr18:74090433-74090815	134.35	91.2%	69.7%	21.4%	hyper
chr14:26577035-26577596	63.27	77.8%	56.4%	21.4%	hyper
chr10:110510234-110510688	69.35	78.5%	57.1%	21.4%	hyper
chr16:13541050-13541709	71.60	80.5%	59.1%	21.4%	hyper
chr4:129807866-129808438	63.64	73.3%	51.9%	21.4%	hyper
chr2:152201019-152201471	66.04	52.6%	31.2%	21.4%	hyper
chr17:81246598-81246934	98.43	84.5%	63.1%	21.4%	hyper
chrX:50251732-50251950	73.79	84.1%	62.7%	21.4%	hyper
chr18:38234553-38235532	94.89	38.7%	17.3%	21.4%	hyper
chr17:24853108-24853646	91.47	91.3%	70.0%	21.3%	hyper
chr9:74856794-74857461	98.09	78.8%	57.5%	21.3%	hyper
chr18:9280454-9280994	83.76	86.1%	64.8%	21.3%	hyper
chr5:112430812-112431751	149.42	63.6%	42.3%	21.3%	hyper
chr19:16021483-16021844	61.27	80.2%	58.9%	21.3%	hyper
chr19:45237509-45238254	118.53	40.8%	19.5%	21.3%	hyper
chr3:107286320-107286787	91.80	81.2%	59.9%	21.3%	hyper
chr1:33720597-33720867	79.31	82.8%	61.5%	21.3%	hyper
chr6:115912422-115912546	74.82	84.0%	62.7%	21.3%	hyper
chr6:30625449-30625743	59.20	67.3%	46.0%	21.3%	hyper

chr18:53620153-53621344	168.87	41.2%	19.9%	21.3%	hyper
chr17:6988252-6988957	63.62	80.4%	59.1%	21.3%	hyper
chr12:107004564-107005251	111.33	80.8%	59.5%	21.3%	hyper
chr5:141088065-141088599	99.91	55.9%	34.7%	21.2%	hyper
chr8:110395832-110396267	90.22	86.0%	64.8%	21.2%	hyper
chr8:111884528-111885110	99.86	75.8%	54.6%	21.2%	hyper
chr6:38876247-38877355	185.69	77.8%	56.6%	21.2%	hyper
chr17:15519420-15519960	65.68	80.6%	59.4%	21.2%	hyper
chr17:48195064-48195640	70.47	71.5%	50.3%	21.2%	hyper
chr5:140275986-140276250	90.29	89.5%	68.3%	21.2%	hyper
chr4:133091530-133092205	85.82	89.7%	68.5%	21.2%	hyper
chr18:10541487-10541931	68.54	90.1%	68.9%	21.2%	hyper
chr4:129825701-129826354	86.66	87.5%	66.4%	21.1%	hyper
chr9:118156308-118156805	72.99	80.4%	59.3%	21.1%	hyper
chr13:53648996-53649858	80.20	71.0%	49.8%	21.1%	hyper
chr9:63000214-63000906	71.32	82.0%	60.9%	21.1%	hyper
chr12:13362825-13363327	64.44	86.8%	65.7%	21.1%	hyper
chr9:106723350-106723628	65.71	70.5%	49.4%	21.1%	hyper
chr19:17639325-17640374	86.05	94.2%	73.1%	21.1%	hyper
chr12:87584914-87585281	91.72	87.6%	66.6%	21.1%	hyper
chr12:104125669-104126130	59.23	81.9%	60.8%	21.1%	hyper
chr2:13051266-13051893	75.30	85.8%	64.7%	21.1%	hyper
chr8:122691960-122692397	60.95	85.4%	64.3%	21.0%	hyper
chr5:125537788-125538345	78.77	44.1%	23.0%	21.0%	hyper
chr1:188243897-188244383	65.06	88.6%	67.6%	21.0%	hyper
chr2:156536576-156537170	151.53	81.0%	60.0%	21.0%	hyper
chr17:5308142-5308630	78.75	85.8%	64.8%	21.0%	hyper
chr6:88154810-88156434	163.82	62.6%	41.6%	21.0%	hyper
chr11:116848981-116849900	71.46	86.0%	65.0%	21.0%	hyper
chr7:143238179-143239028	76.72	70.8%	49.8%	21.0%	hyper
chr16:57473874-57474235	84.08	83.4%	62.4%	21.0%	hyper
chr5:113901210-113901650	59.60	79.2%	58.3%	20.9%	hyper
chr2:46865214-46865441	62.99	78.2%	57.3%	20.9%	hyper
chr7:30884161-30884649	78.88	79.2%	58.4%	20.9%	hyper
chr19:5751116-5751276	55.05	40.8%	19.9%	20.9%	hyper
chr13:38404035-38404691	80.95	84.3%	63.5%	20.9%	hyper
chr2:158222992-158223588	72.03	87.9%	67.0%	20.9%	hyper
chr7:104473728-104473923	101.86	40.5%	19.7%	20.9%	hyper
chr7:151793769-151793980	90.62	90.6%	69.7%	20.9%	hyper
chr18:50289206-50289570	76.90	85.5%	64.7%	20.8%	hyper
chr18:56933833-56934296	91.43	53.0%	32.2%	20.8%	hyper
chr6:102723509-102724005	79.53	85.7%	64.9%	20.8%	hyper
chr19:34749463-34749849	136.60	87.3%	66.5%	20.8%	hyper
chr14:78607638-78608465	82.08	88.4%	67.5%	20.8%	hyper

chr3:138451219-138452302	80.05	79.7%	58.9%	20.8%	hyper
chr12:74468614-74468871	70.93	57.2%	36.4%	20.8%	hyper
chr2:4524706-4524912	64.89	84.4%	63.6%	20.8%	hyper
chr2:56960315-56961087	75.23	55.7%	34.9%	20.8%	hyper
chr2:52614822-52615371	78.52	86.3%	65.5%	20.8%	hyper
chr10:3996841-3997235	67.42	80.9%	60.1%	20.8%	hyper
chr12:74136398-74136874	62.82	27.5%	6.7%	20.8%	hyper
chr7:139439228-139439778	93.39	56.6%	35.8%	20.8%	hyper
chr1:34617333-34617818	81.34	87.8%	67.0%	20.8%	hyper
chr7:117952877-117953693	82.11	86.9%	66.2%	20.8%	hyper
chr2:181687545-181688517	108.20	61.0%	40.2%	20.8%	hyper
chr13:51324601-51324899	76.18	68.0%	47.2%	20.8%	hyper
chr4:150096149-150096619	66.24	78.7%	58.0%	20.8%	hyper
chr9:74866935-74867563	64.64	88.2%	67.4%	20.7%	hyper
chr6:35766730-35767088	95.15	86.6%	65.8%	20.7%	hyper
chr4:107732383-107733085	66.09	79.1%	58.4%	20.7%	hyper
chr1:190479902-190480475	78.33	90.2%	69.5%	20.7%	hyper
chr12:86021533-86021896	83.02	78.0%	57.4%	20.6%	hyper
chr1:94745279-94745786	72.23	47.4%	26.8%	20.6%	hyper
chr19:28240382-28240692	69.82	77.0%	56.4%	20.6%	hyper
chr5:116728472-116728957	79.14	89.9%	69.3%	20.6%	hyper
chr14:80754829-80755288	65.35	84.5%	63.9%	20.6%	hyper
chr2:173876541-173876778	87.42	55.9%	35.3%	20.6%	hyper
chr1:94120717-94121373	97.04	80.6%	60.0%	20.6%	hyper
chr11:99793470-99793616	77.62	78.0%	57.4%	20.6%	hyper
chr4:151685683-151686122	74.61	73.5%	52.9%	20.6%	hyper
chr1:55249035-55249479	71.90	65.1%	44.5%	20.6%	hyper
chr5:34073804-34074166	104.01	78.3%	57.7%	20.6%	hyper
chr18:67632405-67632679	73.84	79.6%	59.1%	20.6%	hyper
chr9:31020271-31021226	78.26	76.7%	56.1%	20.6%	hyper
chr17:31418966-31419653	100.94	59.2%	38.7%	20.6%	hyper
chr19:21218809-21219751	67.19	52.2%	31.6%	20.5%	hyper
chr15:75608662-75609044	81.55	89.4%	68.9%	20.5%	hyper
chr17:7296414-7296733	80.20	87.6%	67.1%	20.5%	hyper
chr11:69234465-69235215	83.79	73.5%	53.0%	20.5%	hyper
chr2:106636573-106637888	63.52	78.7%	58.2%	20.5%	hyper
chr9:6168439-6168894	65.11	58.1%	37.7%	20.5%	hyper
chr19:55662831-55663951	92.40	49.1%	28.6%	20.5%	hyper
chr7:137910559-137911014	68.48	81.1%	60.6%	20.5%	hyper
chr14:120003673-120004158	88.51	81.2%	60.8%	20.5%	hyper
chr1:184002222-184002965	89.19	85.4%	64.9%	20.5%	hyper
chr16:22766850-22767373	78.19	83.9%	63.4%	20.5%	hyper
chr11:4170937-4171826	169.12	58.1%	37.6%	20.5%	hyper
chr9:21578977-21579304	83.94	87.0%	66.6%	20.4%	hyper

chr10:93982857-93983441	65.92	79.6%	59.1%	20.4%	hyper
chr1:88967343-88967536	72.15	73.5%	53.1%	20.4%	hyper
chr5:36656105-36656649	59.01	69.3%	48.8%	20.4%	hyper
chr11:76165957-76166324	61.96	79.5%	59.1%	20.4%	hyper
chr11:105317580-105317884	69.27	41.4%	21.0%	20.4%	hyper
chr9:119392885-119393333	84.51	74.8%	54.5%	20.4%	hyper
chr4:54833930-54834429	80.40	82.0%	61.7%	20.3%	hyper
chr6:4943803-4944395	65.94	76.2%	55.9%	20.3%	hyper
chr3:54921411-54921698	69.29	86.7%	66.4%	20.3%	hyper
chr4:106499486-106500075	75.77	72.0%	51.7%	20.3%	hyper
chr5:115179095-115179494	82.56	85.4%	65.2%	20.3%	hyper
chr11:100819860-100820506	370.50	45.7%	25.5%	20.2%	hyper
chr3:85422589-85423143	70.74	87.1%	66.9%	20.2%	hyper
chr19:37762177-37762475	90.42	52.9%	32.7%	20.2%	hyper
chr7:130228326-130228563	71.50	63.4%	43.2%	20.2%	hyper
chr12:100249122-100249527	68.65	32.5%	12.3%	20.2%	hyper
chr4:131478296-131478516	61.17	62.6%	42.4%	20.1%	hyper
chr4:137037907-137039268	89.64	78.3%	58.2%	20.1%	hyper
chr6:69443602-69444004	91.52	82.9%	62.8%	20.1%	hyper
chr2:152896544-152896824	107.10	43.1%	23.0%	20.1%	hyper
chr3:133028693-133029300	61.33	86.7%	66.6%	20.1%	hyper
chr11:94984140-94984908	87.77	65.4%	45.3%	20.1%	hyper
chr5:38757804-38758370	78.51	49.0%	28.9%	20.1%	hyper
chr10:22682551-22683395	98.52	86.8%	66.7%	20.1%	hyper
chr13:38286975-38287243	59.31	86.2%	66.1%	20.1%	hyper
chr13:31903632-31904069	183.77	48.6%	28.6%	20.1%	hyper
chr1:91822191-91822657	61.02	60.7%	40.7%	20.0%	hyper
chr4:104608426-104609225	72.67	79.5%	59.4%	20.0%	hyper
chr7:105995667-105995984	118.12	89.7%	69.6%	20.0%	hyper
chr3:68558045-68558680	102.49	92.4%	72.4%	20.0%	hyper
chr12:112661520-112662384	282.90	40.5%	20.5%	20.0%	hyper
chr14:43390982-43391331	79.60	77.8%	57.8%	20.0%	hyper
chr4:141920022-141920620	-88.12	68.3%	88.5%	-20.2%	hypo
chr2:26330278-26331358	-107.29	41.9%	62.2%	-20.3%	hypo
chr18:35253226-35254646	-132.46	36.2%	56.6%	-20.3%	hypo
chr9:63766729-63767498	-119.88	53.8%	74.2%	-20.4%	hypo
chr4:45801580-45802815	-138.63	58.7%	79.1%	-20.4%	hypo
chr12:105844170-105844806	-203.50	48.0%	68.4%	-20.4%	hypo
chr8:124820289-124820793	-128.09	30.6%	51.1%	-20.6%	hypo
chr8:123356063-123357085	-71.23	31.1%	51.7%	-20.7%	hypo
chr10:74450441-74451189	-110.64	57.1%	77.8%	-20.7%	hypo
chr13:112658019-112658630	-157.80	54.8%	75.5%	-20.7%	hypo
chr1:143038343-143039131	-113.58	44.5%	65.3%	-20.8%	hypo
chr2:163054748-163055533	-133.32	55.3%	76.3%	-21.0%	hypo

chr11:103662445-103663150	-91.48	60.6%	81.6%	-21.0%	hypo
chr11:84514834-84515530	-142.14	35.6%	56.7%	-21.1%	hypo
chr18:60868700-60869625	-133.94	34.8%	56.0%	-21.1%	hypo
chr19:37814091-37814949	-99.55	56.4%	77.6%	-21.2%	hypo
chr8:124849801-124850807	-90.20	24.3%	45.6%	-21.4%	hypo
chr11:19863627-19864000	-115.18	48.3%	69.8%	-21.5%	hypo
chr2:92500709-92501400	-81.88	39.1%	60.5%	-21.5%	hypo
chr10:19660665-19661177	-105.00	57.8%	79.4%	-21.6%	hypo
chr14:47323625-47324529	-91.32	52.8%	74.5%	-21.7%	hypo
chr11:78646352-78647093	-125.08	44.8%	66.8%	-21.9%	hypo
chr7:150243615-150244503	-96.66	37.8%	59.7%	-21.9%	hypo
chr10:74816185-74817404	-110.76	58.1%	80.1%	-22.0%	hypo
chr8:124845529-124845986	-78.30	35.3%	57.8%	-22.4%	hypo
chr13:45126721-45127513	-128.04	49.8%	72.3%	-22.5%	hypo
chr10:125222250-125223145	-100.75	23.9%	46.8%	-22.9%	hypo
chr19:9553902-9554571	-62.90	53.7%	76.6%	-22.9%	hypo
chr4:134156973-134158045	-98.26	52.0%	75.0%	-23.0%	hypo
chr5:150057323-150058420	-112.98	35.9%	59.0%	-23.1%	hypo
chr5:120465365-120465988	-99.10	47.1%	70.4%	-23.3%	hypo
chr19:36203637-36204826	-160.14	22.1%	46.3%	-24.2%	hypo
chr8:11775411-11776517	-129.38	45.8%	70.1%	-24.3%	hypo
chr8:94041532-94042770	-121.68	49.6%	74.2%	-24.6%	hypo
chr15:73828139-73828384	-72.18	42.8%	68.0%	-25.1%	hypo
chr8:123710976-123711748	-135.34	57.7%	83.3%	-25.6%	hypo
chr6:100521027-100521575	-118.13	35.4%	61.2%	-25.8%	hypo
chr8:118084907-118086074	-130.17	52.8%	79.5%	-26.7%	hypo
chr7:86503189-86504238	-141.57	35.9%	63.3%	-27.5%	hypo
chr9:114899857-114900413	-104.43	29.5%	58.3%	-28.8%	hypo
chr2:34619760-34620350	-81.89	36.6%	65.9%	-29.3%	hypo
chr2:101613620-101614322	-123.00	49.9%	79.5%	-29.6%	hypo
chr11:109292480-109293051	-133.48	40.1%	70.6%	-30.5%	hypo
chr17:6208973-6209578	-107.73	38.8%	70.9%	-32.1%	hypo
chr3:157574510-157574815	-118.13	43.1%	76.9%	-33.8%	hypo

A.4 Differentially Regulated Genes between Endocardium and Endothelium from Embryoid Body cultures

Table A.4: Complete list of differentially regulated genes between the endocardium and the endothelium sorted by Fold Change.

Gene	Locus	EC Expression	ET expression	p-value	q-value	Fold Change
Oit3	chr10:58885707-58904527	5.20	0.30	5.00E-05	2.50E-03	4.10
Slc32a1	chr2:158436493-158441483	5.13	0.31	5.00E-05	2.50E-03	4.05
Gad2	chr2:22477846-22549397	4.57	0.43	5.00E-05	2.50E-03	3.42
Ptprb	chr10:115738429-115826594	2.53	0.25	5.00E-05	2.50E-03	3.31
-	chr15:72333348-72335718	2.09	0.22	5.50E-04	1.78E-02	3.26
Sost	chr11:101823771-101828329	4.49	0.51	5.00E-05	2.50E-03	3.13
Ccm2l	chr2:152891690-152907471	2.85	0.34	1.15E-03	3.19E-02	3.05
Cdh5	chr8:106625524-106668402	37.48	4.59	5.00E-05	2.50E-03	3.03
Sox7	chr14:64562542-64569569	10.68	1.39	5.00E-05	2.50E-03	2.94
Erg	chr16:95581810-95751972	11.47	1.51	5.00E-05	2.50E-03	2.92
Nos3	chr5:23870636-23897961	6.88	0.94	5.00E-05	2.50E-03	2.88
Grap	chr11:61466822-61486279	6.70	0.91	5.00E-05	2.50E-03	2.87
Lyve1	chr7:117994120-118006467	1.68	0.23	1.25E-03	3.41E-02	2.87
Dusp2	chr2:127161894-127164113	7.11	0.98	5.00E-05	2.50E-03	2.86
Rasip1	chr7:52882906-52894462	12.56	1.84	5.00E-05	2.50E-03	2.77
Eltd1	chr3:151100845-151208045	3.66	0.54	5.00E-05	2.50E-03	2.77
Skap1	chr11:96325904-96620936	5.66	0.84	5.00E-05	2.50E-03	2.76
Gimap4	chr6:48634576-48642061	2.02	0.30	2.00E-04	7.94E-03	2.74
Samsn1	chr16:75859038-75909511	4.03	0.62	5.00E-05	2.50E-03	2.70
Ushbp1	chr8:73908172-73919704	2.22	0.35	5.00E-05	2.50E-03	2.66
Tie1	chr4:118143795-118162454	40.45	6.62	5.00E-05	2.50E-03	2.61
Cd93	chr2:148262386-148269271	13.66	2.26	5.00E-05	2.50E-03	2.60
Kcne3	chr7:107325179-107333379	8.80	1.46	5.00E-05	2.50E-03	2.59
-	chr15:72337445-72342362	1.73	0.29	1.00E-04	4.56E-03	2.58
-	chr16:95580786-95581555	10.43	1.76	5.00E-05	2.50E-03	2.57
Adra2a	chr19:54119671-54123472	3.92	0.67	5.00E-05	2.50E-03	2.56
Icam2	chr11:106238969-106243955	13.43	2.29	5.00E-05	2.50E-03	2.55
Gimap6	chr6:48651581-48658243	6.43	1.11	5.00E-05	2.50E-03	2.54
Emcn	chr3:137004041-137094033	7.16	1.24	5.00E-05	2.50E-03	2.53
Ikzf1	chr11:11586215-11672929	4.18	0.74	5.00E-05	2.50E-03	2.50
Esam	chr9:37335673-37345904	3.97	0.72	1.70E-03	4.33E-02	2.47
Egfl7	chr2:26436575-26448202	12.09	2.19	5.00E-05	2.50E-03	2.47
Myct1	chr10:4739751-4752813	2.78	0.51	1.50E-04	6.36E-03	2.46
Myzap	chr9:71352153-71440167	9.76	1.79	5.00E-05	2.50E-03	2.45
Esam	chr9:37335673-37345904	27.65	5.10	5.00E-05	2.50E-03	2.44
Tspan8	chr10:115254339-115286949	2.12	0.39	7.00E-04	2.15E-02	2.43
Gfi1b	chr2:28464969-28477502	8.33	1.54	5.00E-05	2.50E-03	2.43
Abi3	chr11:95685813-95707045	4.32	0.83	1.00E-04	4.56E-03	2.38
Pcdh12	chr18:38426745-38444055	3.60	0.70	5.00E-05	2.50E-03	2.36
Sox18	chr2:181404541-181406345	4.31	0.85	1.50E-04	6.36E-03	2.34

Csgalnact1	chr8:70880679-71259045	4.68	0.93	5.00E-05	2.50E-03	2.34
Adora2a	chr10:74779687-74797533	4.71	0.93	5.00E-05	2.50E-03	2.34
Mmrn2	chr14:35188689-35217472	12.11	2.41	5.00E-05	2.50E-03	2.33
Acer2	chr4:86520317-86566785	13.32	2.66	5.00E-05	2.50E-03	2.32
Thsd1	chr8:23337774-23371804	4.92	0.99	5.00E-05	2.50E-03	2.32
Tal1	chr4:114729365-114744360	31.18	6.31	5.00E-05	2.50E-03	2.31
Mpo	chr11:87607285-87617914	2.37	0.48	5.50E-04	1.78E-02	2.30
Afap1l1	chr18:61889053-61946316	17.92	3.69	5.00E-05	2.50E-03	2.28
Gngt2	chr11:95685813-95707045	21.37	4.46	1.75E-03	4.41E-02	2.26
Igf1	chr10:87321800-87399792	8.47	1.78	5.00E-05	2.50E-03	2.25
Tal1	chr4:114729365-114744360	5.76	1.21	1.15E-03	3.19E-02	2.25
Pde2a	chr7:108570204-108661343	8.98	1.90	5.00E-05	2.50E-03	2.24
Bcl6b	chr11:70037628-70043300	4.24	0.90	5.00E-05	2.50E-03	2.23
Tal1	chr4:114729365-114744360	6.20	1.32	6.50E-04	2.05E-02	2.23
Robo4	chr9:37209630-37221607	3.30	0.72	5.00E-05	2.50E-03	2.20
Reln	chr5:21390271-21850523	9.72	2.15	5.00E-05	2.50E-03	2.17
Cav2	chr6:17231340-17239011	1.81	0.40	6.00E-04	1.93E-02	2.17
Car8	chr4:8068639-8166188	2.54	0.57	5.00E-05	2.50E-03	2.16
Hapln1	chr13:89680240-89751437	42.57	9.67	5.00E-05	2.50E-03	2.14
-	chr18:38158929-38159188	80.70	18.33	1.00E-04	4.56E-03	2.14
Lmo2	chr2:103798151-103822035	26.72	6.16	5.00E-05	2.50E-03	2.12
Fxyd5	chr7:31817741-31827341	8.40	1.94	1.50E-04	6.36E-03	2.12
Scarf1	chr11:75327042-75340082	9.77	2.27	5.00E-05	2.50E-03	2.10
Stab1	chr14:31944314-31981827	15.70	3.68	5.00E-05	2.50E-03	2.09
Nfatc1	chr18:80802943-80909810	11.84	2.78	5.00E-05	2.50E-03	2.09
C1qc	chr4:136445716-136448829	7.76	1.84	9.00E-04	2.63E-02	2.07
Flt4	chr11:49423180-49466241	14.91	3.56	5.00E-05	2.50E-03	2.07
Rasgrp3	chr17:75835244-75928393	11.17	2.72	5.00E-05	2.50E-03	2.04
Rhoj	chr12:76409299-76502442	8.84	2.16	5.00E-05	2.50E-03	2.04
Gata2	chr6:88148657-88157026	8.45	2.07	5.00E-05	2.50E-03	2.03
Ctla2b	chr13:60996711-60998808	7.13	1.76	8.50E-04	2.53E-02	2.02
Fli1	chr9:32229792-32349149	8.58	2.14	5.00E-05	2.50E-03	2.00
Gmfg	chr7:29222465-29231914	15.05	3.75	1.50E-04	6.36E-03	2.00
Afp	chr5:90919739-90937933	3.51	0.88	5.50E-04	1.78E-02	2.00
Plvap	chr8:74021651-74035668	19.45	4.98	5.00E-05	2.50E-03	1.96
She	chr3:89635291-89662768	3.96	1.03	5.00E-05	2.50E-03	1.94
Gja4	chr4:126988663-126991283	9.52	2.49	5.00E-05	2.50E-03	1.93
Cav1	chr6:17256334-17291328	3.82	1.01	1.50E-04	6.36E-03	1.92
Hhex	chr19:37509330-37515221	15.41	4.11	5.00E-05	2.50E-03	1.91
Dll4	chr2:119146934-119161402	4.04	1.08	5.00E-05	2.50E-03	1.90
Wnk4	chr11:101121880-101140262	11.85	3.24	5.00E-05	2.50E-03	1.87
Gng11	chr6:3953986-3958445	13.02	3.58	1.50E-04	6.36E-03	1.86
Mrc1	chr2:14151040-14253650	7.45	2.08	5.00E-05	2.50E-03	1.84
Spns2	chr11:72265139-72303422	13.03	3.64	5.00E-05	2.50E-03	1.84

Asic4	chr1:75447084-75470915	5.47	1.54	5.00E-05	2.50E-03	1.83
C430049B03Rik, Mir322, Mir351, Mir503	chrX:50406288-50410367	33.35	9.37	5.00E-05	2.50E-03	1.83
Alox5ap	chr5:150076633-150099623	13.46	3.78	1.50E-04	6.36E-03	1.83
Fgf3	chr7:152024516-152030660	14.53	4.09	5.00E-05	2.50E-03	1.83
Sema3g	chr14:32031058-32042697	1.80	0.51	5.50E-04	1.78E-02	1.83
Kcnj5	chr9:32122367-32151822	2.19	0.62	5.00E-05	2.50E-03	1.82
Ctla2a	chr13:61035515-61037986	23.45	6.68	5.00E-05	2.50E-03	1.81
Gpm6b	chrX:162676874-162826965	10.21	2.91	5.00E-05	2.50E-03	1.81
Emr1	chr17:57498108-57622952	2.88	0.83	4.00E-04	1.40E-02	1.80
Ramp2	chr11:101107647-101109564	49.56	14.33	5.00E-05	2.50E-03	1.79
Fam212a	chr9:107886554-107888247	23.05	6.77	5.00E-05	2.50E-03	1.77
Cldn5	chr16:18776939-18778355	9.86	2.92	5.00E-05	2.50E-03	1.76
Ccl9	chr11:83386418-83392138	4.27	1.26	5.00E-04	1.67E-02	1.76
Eng	chr2:32502114-32549695	17.93	5.32	5.00E-05	2.50E-03	1.75
Notch4	chr17:34701239-34725488	3.50	1.04	5.00E-05	2.50E-03	1.75
Aqp1	chr6:55286292-55298549	2.07	0.62	1.65E-03	4.22E-02	1.74
Myb	chr10:20844735-20880790	6.65	2.01	5.00E-05	2.50E-03	1.73
Arhgef15	chr11:68756655-68770424	5.36	1.62	5.00E-05	2.50E-03	1.72
Tnfaip2	chr12:112680871-112693229	24.62	7.49	5.00E-05	2.50E-03	1.72
Lmo2	chr2:103798151-103822035	3.42	1.04	1.30E-03	3.52E-02	1.72
Gp49a	chr10:51200484-51206025	2.61	0.80	1.00E-03	2.86E-02	1.71
Lyl1	chr8:87225355-87229783	7.54	2.30	1.50E-04	6.36E-03	1.71
Cd34	chr1:196765014-196803153	28.63	8.75	5.00E-05	2.50E-03	1.71
Sox17	chr1:4481008-4486494	2.22	0.68	1.35E-03	3.63E-02	1.70
S1pr1	chr3:115413350-115417973	7.94	2.45	5.00E-05	2.50E-03	1.69
Slc18a2	chr19:59335367-59370502	16.84	5.23	5.00E-05	2.50E-03	1.69
Plxnd1	chr6:115904828-115945023	42.55	13.29	5.00E-05	2.50E-03	1.68
Exoc3l	chr8:107813823-107820136	4.72	1.48	1.00E-04	4.56E-03	1.68
Ctse	chr1:133534890-133572084	7.22	2.26	5.00E-05	2.50E-03	1.67
Cd38	chr5:44260065-44303613	5.10	1.61	5.00E-05	2.50E-03	1.66
P2rx1	chr11:72812646-72828699	9.37	2.97	5.00E-05	2.50E-03	1.66
Cbfa2t3	chr8:125149035-125223009	5.00	1.60	5.00E-05	2.50E-03	1.64
Prkch	chr12:74686027-74879171	3.08	0.99	2.00E-04	7.94E-03	1.64
Lgr5	chr10:114887369-115024836	2.48	0.79	1.50E-04	6.36E-03	1.64
Calcrl	chr2:84170782-84265423	11.18	3.59	5.00E-05	2.50E-03	1.64
Vav3	chr3:109143600-109488612	7.50	2.41	5.00E-05	2.50E-03	1.64
F10	chr8:13037307-13056676	5.10	1.64	2.50E-04	9.49E-03	1.63
Fgf3	chr7:152024516-152030660	17.22	5.56	1.00E-04	4.56E-03	1.63
Fcgr3	chr1:172981299-172989534	5.17	1.68	1.90E-03	4.71E-02	1.62
C3ar1	chr6:122797157-122806175	2.62	0.86	7.50E-04	2.27E-02	1.60
Fam198b	chr3:79689851-79750200	2.95	0.97	5.00E-05	2.50E-03	1.60
Adcy4	chr14:56387928-56402856	3.60	1.19	5.00E-05	2.50E-03	1.60
C1qb	chr4:136436060-136442092	11.57	3.85	7.00E-04	2.15E-02	1.59
Ets2	chr16:95924013-95942656	28.51	9.59	5.00E-05	2.50E-03	1.57

Lilrb4	chr10:51210780-51216417	5.39	1.81	5.00E-05	2.50E-03	1.57
Klhl4	chrX:111588128-111674307	4.11	1.39	5.00E-05	2.50E-03	1.57
Enpp1	chr10:24361216-24431908	5.02	1.70	5.00E-05	2.50E-03	1.56
Cyth4	chr15:78427476-78452449	4.13	1.40	6.50E-04	2.05E-02	1.56
Tyrobp	chr7:31198806-31202598	30.08	10.23	2.00E-04	7.94E-03	1.56
Gata1	chrX:7536385-7545036	9.11	3.12	5.00E-05	2.50E-03	1.55
Klhl6	chr16:19946585-19983122	5.38	1.86	5.00E-04	1.67E-02	1.53
Clic6	chr16:92498391-92541486	9.27	3.22	5.00E-05	2.50E-03	1.53
-	chr7:20080256-20082313	5.02	1.75	4.50E-04	1.54E-02	1.52
Tmem255a	chrX:35550477-35605658	1.71	0.60	2.00E-03	4.91E-02	1.52
Nfe2	chr15:103078648-103085847	7.17	2.56	5.00E-04	1.67E-02	1.48
Col15a1	chr4:47220883-47326037	1.66	0.59	8.50E-04	2.53E-02	1.48
Rapgef3	chr15:97575200-97598097	2.75	1.00	6.50E-04	2.05E-02	1.46
Runx1	chr16:92601710-92826311	5.95	2.17	5.00E-05	2.50E-03	1.45
Nckap1l	chr15:103284255-103329231	2.56	0.94	5.50E-04	1.78E-02	1.45
Etv2	chr7:31418634-31421103	13.03	4.80	5.50E-04	1.78E-02	1.44
Myo1f	chr17:33692651-33744709	2.70	1.00	1.40E-03	3.74E-02	1.44
Arap3	chr18:38132276-38158623	27.01	9.96	5.00E-05	2.50E-03	1.44
Ptpre	chr7:142729506-142877977	7.29	2.69	5.00E-05	2.50E-03	1.44
Padi3	chr4:140341283-140366563	11.43	4.26	5.00E-05	2.50E-03	1.42
Shank3	chr15:89330287-89390691	5.83	2.19	5.00E-05	2.50E-03	1.41
Trf	chr9:103111205-103132616	3.96	1.50	1.30E-03	3.52E-02	1.40
Stat5a	chr11:100720664-100746483	4.92	1.87	5.00E-05	2.50E-03	1.40
Aplnr	chr2:84976516-84980080	53.46	20.67	5.00E-05	2.50E-03	1.37
Csf1r	chr18:61265225-61290793	10.79	4.20	5.00E-05	2.50E-03	1.36
Car2	chr3:14886425-14900770	39.24	15.35	5.00E-05	2.50E-03	1.35
Fgd5	chr6:91937103-92025999	12.15	4.77	5.00E-05	2.50E-03	1.35
Ubash3b	chr9:40819212-40965577	7.64	3.00	5.00E-05	2.50E-03	1.35
Rspo3	chr10:29166747-29255673	32.90	12.93	5.00E-05	2.50E-03	1.35
Edil3	chr13:88961076-89462830	9.67	3.81	5.00E-05	2.50E-03	1.35
Tiam1	chr16:89787355-89974944	18.79	7.41	5.00E-05	2.50E-03	1.34
Hid1	chr11:115209022-115229033	3.61	1.42	7.50E-04	2.27E-02	1.34
Plxnc1	chr10:94253611-94407212	5.73	2.26	5.00E-05	2.50E-03	1.34
Atp2a3	chr11:72774670-72806545	22.50	8.95	5.00E-05	2.50E-03	1.33
Flt1	chr5:148373771-148537564	10.32	4.11	3.50E-04	1.26E-02	1.33
Fam78a	chr2:31922404-31939225	3.42	1.38	3.50E-04	1.26E-02	1.31
Cped1	chr6:21935909-22205606	2.87	1.15	3.00E-04	1.11E-02	1.31
Sepp1	chr15:3220766-3230508	61.77	24.94	5.00E-05	2.50E-03	1.31
Slc30a10	chr1:187270994-187292640	4.53	1.83	5.00E-05	2.50E-03	1.31
Kdr	chr5:76329298-76374453	113.86	46.03	5.00E-05	2.50E-03	1.31
Pdgfb	chr15:79826305-79845238	8.94	3.63	1.00E-04	4.56E-03	1.30
Cx3cr1	chr9:119957800-119977414	11.84	4.81	5.00E-05	2.50E-03	1.30
Lrrc33	chr16:32142910-32165562	4.36	1.78	3.00E-04	1.11E-02	1.29
Tgm2	chr2:157942140-157972128	13.27	5.47	5.00E-05	2.50E-03	1.28

Prkar2b	chr12:32643343-32746144	37.23	15.50	5.00E-05	2.50E-03	1.26
Map4k2	chr19:6341249-6356180	13.10	5.47	5.00E-05	2.50E-03	1.26
Ptprc	chr1:139959435-140071882	2.51	1.05	1.05E-03	2.98E-02	1.26
Egfl7	chr2:26436575-26448202	74.37	31.21	5.00E-05	2.50E-03	1.25
Epor	chr9:21763342-21768020	11.37	4.80	4.50E-04	1.54E-02	1.25
Shank3	chr15:89330287-89390691	8.22	3.53	5.00E-05	2.50E-03	1.22
Myl7	chr11:5796639-5798785	50.63	22.05	1.00E-04	4.56E-03	1.20
Flt1	chr5:148373771-148537564	17.26	7.59	5.00E-05	2.50E-03	1.18
Arhgap18	chr10:26492317-26638454	6.39	2.82	5.00E-05	2.50E-03	1.18
Gpr116	chr17:43526414-43596506	4.36	1.93	5.00E-05	2.50E-03	1.18
Lcp1	chr14:75530929-75630649	7.34	3.26	1.00E-04	4.56E-03	1.17
Laptm5	chr4:130469248-130492063	20.37	9.05	5.00E-05	2.50E-03	1.17
Unc13d	chr11:115923409-115939275	4.32	1.93	5.00E-04	1.67E-02	1.16
Rapgef5	chr12:118754951-118998177	8.03	3.60	1.35E-03	3.63E-02	1.16
Slc7a8	chr14:55341051-55400723	4.96	2.23	4.50E-04	1.54E-02	1.15
Pear1	chr3:87553018-87572875	13.00	5.88	5.00E-05	2.50E-03	1.14
Dock8	chr19:25074018-25276922	2.22	1.01	3.50E-04	1.26E-02	1.14
Thbd	chr2:148230206-148233924	2.88	1.31	1.60E-03	4.12E-02	1.14
Map4k2	chr19:6341249-6356180	20.97	9.53	5.00E-05	2.50E-03	1.14
Zfp711	chrX:109714134-109748671	4.05	1.84	2.00E-04	7.94E-03	1.14
Adamtsl2	chr2:26934900-26964133	7.79	3.54	2.00E-04	7.94E-03	1.14
Cyp26b1	chr6:84521407-84543902	3.51	1.61	9.00E-04	2.63E-02	1.12
Ecscr	chr18:35872742-35881145	27.04	12.51	1.50E-04	6.36E-03	1.11
Klhl6	chr16:19946585-19983122	29.47	13.65	5.00E-05	2.50E-03	1.11
Elk3	chr10:92710160-92773904	18.79	8.73	5.00E-05	2.50E-03	1.11
Mef2c	chr13:83643032-83806684	7.42	3.45	5.00E-05	2.50E-03	1.11
Sh3tc1	chr5:36039828-36071925	3.92	1.83	6.50E-04	2.05E-02	1.10
Exoc6	chr19:37624907-37758502	8.05	3.77	2.50E-04	9.49E-03	1.10
Rnd2	chr11:101329651-101333780	231.55	109.41	5.00E-05	2.50E-03	1.08
Ets1	chr9:32503626-32565405	32.05	15.19	5.00E-05	2.50E-03	1.08
Capn5	chr7:105270074-105333309	15.25	7.23	5.00E-05	2.50E-03	1.08
Acvrl1	chr15:100958967-100975767	19.44	9.29	5.00E-05	2.50E-03	1.07
Fam43a	chr16:30599808-30602883	7.78	3.72	2.50E-04	9.49E-03	1.07
Myo7a	chr7:105199563-105268003	3.38	1.62	2.00E-04	7.94E-03	1.06
Actc1	chr2:113873024-113878547	29.34	14.07	5.00E-05	2.50E-03	1.06
Sh2b3	chr5:122265469-122286810	17.13	8.22	4.00E-04	1.40E-02	1.06
Bmp2	chr2:133377982-133388621	18.03	8.75	5.00E-05	2.50E-03	1.04
Tspan12	chr6:21721394-21802515	21.86	10.67	5.00E-05	2.50E-03	1.03
St8sia1	chr6:142762750-142912972	2.09	1.02	9.00E-04	2.63E-02	1.03
Zfp711	chrX:109714134-109748671	18.31	8.96	5.00E-05	2.50E-03	1.03
Zfpm1	chr8:124806040-124861147	28.72	14.06	5.00E-05	2.50E-03	1.03
Slc39a8	chr3:135488454-135551536	18.65	9.15	5.00E-05	2.50E-03	1.03
Stat5b	chr11:100642044-100711899	15.20	7.51	5.00E-05	2.50E-03	1.02
Ppp1r16b	chr2:158492468-158592070	3.22	1.60	8.50E-04	2.53E-02	1.01

Grap2	chr15:80453912-80483450	10.22	5.09	8.00E-04	2.41E-02	1.01
Tnnc1	chr14:32021497-32024897	58.41	29.17	1.50E-04	6.36E-03	1.00
Abca1	chr4:53043660-53172767	1.85	0.93	1.00E-03	2.86E-02	1.00
Mpeg1	chr19:12535268-12539775	5.05	2.58	1.90E-03	4.71E-02	0.97
Hspa12a	chr19:58870240-58935474	5.16	2.64	1.00E-04	4.56E-03	0.97
Kctd12	chr14:103375797-103381854	5.78	2.98	2.00E-04	7.94E-03	0.96
Insl3, Jak3	chr8:74200281-74214476	14.58	7.56	5.00E-05	2.50E-03	0.95
Rcsd1	chr1:167579075-167638225	20.12	10.46	5.00E-05	2.50E-03	0.94
Elmo1	chr13:20182375-20700222	3.40	1.80	1.10E-03	3.09E-02	0.92
St3gal1	chr15:66934436-67008444	13.70	7.32	5.00E-05	2.50E-03	0.90
Fam65a	chr8:108129128-108146118	37.58	20.41	8.50E-04	2.53E-02	0.88
Pecam1	chr11:106515531-106585695	52.41	28.46	5.00E-05	2.50E-03	0.88
Mtus1	chr8:42076265-42219080	5.64	3.07	9.00E-04	2.63E-02	0.88
Ddah1	chr3:145421655-145557241	9.42	5.14	2.00E-04	7.94E-03	0.87
Itga9	chr9:118515826-118810121	14.29	7.81	5.00E-05	2.50E-03	0.87
Cd109	chr9:78463352-78564067	4.66	2.56	8.00E-04	2.41E-02	0.86
Lrrc8b	chr5:105844793-105919065	8.42	4.64	1.05E-03	2.98E-02	0.86
Gfod1	chr13:43290887-43399541	5.89	3.25	1.00E-04	4.56E-03	0.86
Cyp26a1	chr19:37772297-37776026	19.22	10.68	2.50E-04	9.49E-03	0.85
Sh2d3c	chr2:32576574-32610527	9.35	5.20	9.00E-04	2.63E-02	0.85
Ptprm	chr17:67016187-67703799	7.79	4.35	1.00E-04	4.56E-03	0.84
Gypc	chr18:32687973-32719688	12.88	7.20	1.20E-03	3.30E-02	0.84
Hoxb3	chr11:96184439-96210872	16.92	9.53	1.00E-04	4.56E-03	0.83
Phlda1	chr10:110943341-110945705	18.76	10.62	5.50E-04	1.78E-02	0.82
Gcnt1	chr19:17400630-17447334	5.63	3.23	1.35E-03	3.63E-02	0.80
Rassf2	chr2:131818585-131855724	6.65	3.82	7.00E-04	2.15E-02	0.80
Bgn	chrX:70728973-70741275	22.34	12.91	1.50E-04	6.36E-03	0.79
Gab2	chr7:104230260-104457461	5.14	2.97	1.05E-03	2.98E-02	0.79
Tnnt2	chr1:137732910-137748845	56.65	32.88	2.50E-04	9.49E-03	0.79
Col23a1	chr11:51103421-51397427	10.51	6.19	5.00E-05	2.50E-03	0.76
Dysf	chr6:83958583-84161036	11.99	7.13	5.00E-05	2.50E-03	0.75
Cap2	chr13:46597271-46745650	9.08	5.40	1.15E-03	3.19E-02	0.75
Taok3	chr5:117570137-117725107	9.12	5.43	1.15E-03	3.19E-02	0.75
Plcl2	chr17:50648871-50827819	7.69	4.61	1.50E-03	3.92E-02	0.74
Dgkz	chr2:91772978-91803720	50.56	30.34	5.00E-05	2.50E-03	0.74
Ppp1r13b	chr12:113066668-113146266	14.51	8.75	5.00E-05	2.50E-03	0.73
Arhgap31	chr16:38598455-38713148	8.98	5.44	3.00E-04	1.11E-02	0.72
Ehbp1l1	chr19:5707373-5726317	10.02	6.10	1.10E-03	3.09E-02	0.72
Gpatch8, Itga2b	chr11:102314610-102417472	60.21	36.80	7.00E-04	2.15E-02	0.71
Hdac7	chr15:97614795-97674933	38.74	23.81	1.00E-04	4.56E-03	0.70
Rasa3	chr8:13567217-13677587	23.38	14.45	5.00E-05	2.50E-03	0.69
Prcp	chr7:100023762-100083091	25.59	15.89	2.50E-04	9.49E-03	0.69
Eln	chr5:135178465-135223124	12.58	7.82	1.55E-03	4.03E-02	0.69
Foxh1	chr15:76498736-76500303	70.89	44.41	2.00E-04	7.94E-03	0.67

Gfra2	chr14:71289936-71379645	11.85	7.47	1.50E-03	3.92E-02	0.66
Plk2	chr13:111185251-111191051	25.16	15.88	5.00E-04	1.67E-02	0.66
Msn	chrX:93291383-93363892	146.07	92.20	1.00E-04	4.56E-03	0.66
Polg	chr7:86537223-86611159	83.43	52.70	1.00E-04	4.56E-03	0.66
Stat3	chr11:100748123-100800825	23.79	15.08	1.25E-03	3.41E-02	0.66
Fzd10	chr5:129106980-129109968	13.82	8.78	1.85E-03	4.62E-02	0.65
Abcb10	chr8:126476358-126507022	23.67	15.08	2.00E-04	7.94E-03	0.65
St6galnac3	chr3:152865472-153388097	12.63	8.05	9.50E-04	2.74E-02	0.65
Myc	chr15:61816895-61821916	18.56	11.86	2.00E-03	4.91E-02	0.65
Map3k11	chr19:5689130-5702864	21.06	13.52	1.10E-03	3.09E-02	0.64
Sipa1	chr19:5651184-5663707	15.84	10.17	1.80E-03	4.51E-02	0.64
Pitpnm2	chr5:124568698-124666427	13.13	8.43	5.00E-05	2.50E-03	0.64
Lama4	chr10:38685320-38829994	30.63	19.68	5.00E-05	2.50E-03	0.64
Rap1b	chr10:117251652-117283030	93.60	60.33	4.50E-04	1.54E-02	0.63
Adam15	chr3:89143561-89153932	19.09	12.36	2.00E-03	4.91E-02	0.63
F2r	chr13:96371743-96388388	210.88	137.19	1.00E-04	4.56E-03	0.62
Ablim1	chr19:57107753-57290522	15.33	10.01	3.50E-04	1.26E-02	0.61
Rreb1	chr13:37917906-38043929	11.15	7.28	4.00E-04	1.40E-02	0.61
Akap5	chr12:77425877-77435138	7.41	4.84	2.00E-03	4.91E-02	0.61
Pros1	chr16:62854159-62929166	27.42	18.01	7.50E-04	2.27E-02	0.61
Cited2	chr10:17443033-17445480	32.47	21.33	1.15E-03	3.19E-02	0.61
Fth1	chr19:10055089-10059601	394.95	260.12	2.00E-04	7.94E-03	0.60
Tgfb1	chr7:26472020-26490015	41.90	27.61	1.30E-03	3.52E-02	0.60
Tmem164	chrX:139115942-139278037	33.81	22.36	4.00E-04	1.40E-02	0.60
Rab11a	chr9:64563106-64585563	33.45	22.17	1.85E-03	4.62E-02	0.59
Vat1	chr11:101320061-101327513	97.84	65.00	2.50E-04	9.49E-03	0.59
Arrb1	chr7:106683995-106755281	22.70	15.08	1.50E-04	6.36E-03	0.59
Prdx3	chr19:60939968-60950441	47.19	31.46	1.75E-03	4.41E-02	0.59
Smad6	chr9:63800882-63869866	22.69	15.14	1.45E-03	3.82E-02	0.58
Ptrf	chr11:100818050-100831931	26.92	18.03	1.45E-03	3.82E-02	0.58
Cycs	chr6:50512561-50516473	30.70	20.60	7.00E-04	2.15E-02	0.58
Epb4.1	chr4:131477064-131631228	37.69	25.74	6.50E-04	2.05E-02	0.55
Ddah2	chr17:35195979-35199044	158.03	108.21	7.50E-04	2.27E-02	0.55
Cpox	chr16:58670320-58680502	29.84	20.43	1.90E-03	4.71E-02	0.55
Sdcbp	chr4:6292826-6323269	53.07	36.38	1.60E-03	4.12E-02	0.54
S100a10	chr3:93359038-93368567	279.31	192.31	1.65E-03	4.22E-02	0.54
Asb4	chr6:5333385-5383021	44.71	31.11	1.70E-03	4.33E-02	0.52
Apoe	chr7:20281592-20284515	422.02	294.88	1.45E-03	3.82E-02	0.52
Slc25a5	chrX:34335646-34338801	200.40	141.61	1.45E-03	3.82E-02	0.50
Ezr	chr17:6942479-6987129	55.86	79.12	1.65E-03	4.22E-02	0.50
Sulf1	chr1:12682510-12850453	52.53	74.66	1.50E-03	3.92E-02	0.51
Slit2	chr5:48374393-48697017	19.12	27.22	1.10E-03	3.09E-02	0.51
Fat1	chr8:46020611-46137611	22.46	32.66	7.00E-04	2.15E-02	0.54
Tmem132c	chr5:127722195-128046160	9.49	14.00	1.45E-03	3.82E-02	0.56

Spon1	chr7:120909511-121186889	11.86	17.50	1.20E-03	3.30E-02	0.56
Lama5	chr2:179911077-179960564	22.43	33.20	3.50E-04	1.26E-02	0.57
Atp6v0a4, D630045J12Rik	chr6:37998482-38204009	4.41	6.54	1.45E-03	3.82E-02	0.57
Axl	chr7:26541518-26573752	14.99	22.37	1.20E-03	3.30E-02	0.58
Ephb2	chr4:136203513-136391850	21.67	32.36	1.15E-03	3.19E-02	0.58
Sall2	chr14:52930851-52948345	11.61	17.42	9.00E-04	2.63E-02	0.59
Bcam	chr7:20341486-20355881	32.47	48.88	4.50E-04	1.54E-02	0.59
Cxcl12	chr6:117118552-117131386	39.00	58.83	1.30E-03	3.52E-02	0.59
Kif5c	chr2:49474833-49630298	8.06	12.30	4.50E-04	1.54E-02	0.61
Nav2	chr7:56214442-56865458	3.35	5.13	1.80E-03	4.51E-02	0.61
Tinagl1	chr4:129842843-129852366	27.21	41.91	8.50E-04	2.53E-02	0.62
Tenm3	chr8:49311018-49760044	16.79	25.92	1.50E-04	6.36E-03	0.63
Sulf2	chr2:165899398-165981183	50.42	77.94	5.00E-05	2.50E-03	0.63
Dsg2	chr18:20716616-20763027	7.66	11.89	9.00E-04	2.63E-02	0.63
Col11a1	chr3:113733457-113923244	16.74	26.04	2.50E-04	9.49E-03	0.64
Cep170b	chr12:113960384-113984802	6.79	10.60	5.50E-04	1.78E-02	0.64
Snhg11	chr2:158201373-158211881	13.70	21.40	1.50E-04	6.36E-03	0.64
Ifi30	chr8:73286671-73291611	38.63	60.68	2.00E-03	4.91E-02	0.65
Emb	chr13:118009379-118063222	44.84	70.69	5.00E-05	2.50E-03	0.66
Cxadr	chr16:78301915-78360030	14.18	22.46	2.50E-04	9.49E-03	0.66
Dlk1	chr12:110691032-110701546	87.66	138.97	4.00E-04	1.40E-02	0.66
Ogdhl	chr14:33135204-33161006	10.68	17.12	2.50E-04	9.49E-03	0.68
Frem1	chr4:82543823-82698071	10.44	16.84	5.00E-05	2.50E-03	0.69
Sesn3	chr9:14080744-14137524	6.57	10.60	5.50E-04	1.78E-02	0.69
Ptprk	chr10:27794625-28317203	11.82	19.20	1.00E-04	4.56E-03	0.70
Ptch1	chr13:63609640-63666828	29.34	47.88	1.10E-03	3.09E-02	0.71
Rpl13a	chr7:52380932-52384115	819.29	1341.58	2.00E-04	7.94E-03	0.71
Dst	chr1:34068669-34365497	9.80	16.06	5.00E-05	2.50E-03	0.71
Ckb	chr12:112907565-112910549	36.01	59.46	5.00E-05	2.50E-03	0.72
Lrrc16b	chr14:56109929-56127101	4.15	6.87	1.90E-03	4.71E-02	0.73
Cacna1h	chr17:25507497-25570728	3.84	6.38	5.50E-04	1.78E-02	0.73
Plat	chr8:23868215-23893320	17.99	30.03	1.00E-04	4.56E-03	0.74
Tenm4	chr7:103359146-104059589	9.33	15.58	2.00E-04	7.94E-03	0.74
H2-Q4	chr17:35516561-35521619	10.87	18.21	1.75E-03	4.41E-02	0.74
Tet1	chr10:62267317-62342762	4.11	6.92	5.00E-05	2.50E-03	0.75
Col8a2	chr4:125964037-125991574	8.90	15.02	5.00E-05	2.50E-03	0.76
Gpc4	chrX:49403563-49518100	10.37	17.60	1.45E-03	3.82E-02	0.76
Shroom3	chr5:93112460-93394785	6.34	10.78	5.00E-05	2.50E-03	0.77
Adamts15	chr9:30706739-30730037	10.68	18.20	5.00E-05	2.50E-03	0.77
Cadm1	chr9:47338434-47661468	21.86	37.48	5.00E-05	2.50E-03	0.78
Sall1	chr8:91551142-91568061	5.43	9.32	5.00E-05	2.50E-03	0.78
Col5a2	chr1:45431175-45560127	31.08	53.48	5.00E-05	2.50E-03	0.78
Skida1	chr2:17965713-17970076	3.51	6.09	1.50E-03	3.92E-02	0.80
Igdcc4	chr9:64949301-64985750	8.44	14.70	5.00E-05	2.50E-03	0.80

Igdcc3	chr9:64988995-65034829	31.48	54.86	5.00E-05	2.50E-03	0.80
Stx3	chr19:11849607-11894262	6.22	10.87	1.80E-03	4.51E-02	0.80
Cdh3	chr8:109034790-109080808	50.76	89.21	5.00E-05	2.50E-03	0.81
Plxnb1	chr9:108997249-109022429	10.67	18.83	5.00E-05	2.50E-03	0.82
Sema4b	chr7:87331726-87371410	5.86	10.35	2.00E-04	7.94E-03	0.82
Arvcf	chr16:18348274-18479166	8.77	15.54	9.50E-04	2.74E-02	0.83
B4galnt4	chr7:148247172-148258018	8.56	15.26	2.50E-04	9.49E-03	0.83
Ptprn2	chr12:117724192-118575485	2.72	4.89	1.75E-03	4.41E-02	0.84
Ptgs2	chr1:151947253-151955142	4.18	7.51	5.50E-04	1.78E-02	0.85
-	chr8:124783835-124796514	1.80	3.24	2.00E-04	7.94E-03	0.85
Pdpn	chr4:142857324-142889410	7.57	13.75	1.45E-03	3.82E-02	0.86
Abca4	chr3:121747377-121882979	1.80	3.28	1.20E-03	3.30E-02	0.87
Fras1	chr5:96802973-97213747	2.85	5.21	5.00E-05	2.50E-03	0.87
Msi1	chr5:115879693-115905695	23.36	42.97	1.25E-03	3.41E-02	0.88
Postn	chr3:54165028-54194963	316.53	583.45	7.50E-04	2.27E-02	0.88
Tnfrsf19	chr14:61582670-61665692	9.47	17.47	5.00E-05	2.50E-03	0.88
Ifitm1	chr7:148153327-148155726	49.99	92.19	5.00E-05	2.50E-03	0.88
Flrt3	chr2:140221165-142215786	8.15	15.09	5.00E-05	2.50E-03	0.89
Cyfip2	chr11:46007350-46126361	2.60	4.82	2.50E-04	9.49E-03	0.89
2610203C20Rik	chr9:41389421-41400570	2.05	3.82	1.75E-03	4.41E-02	0.90
Crabp1	chr9:54612614-54620916	21.95	40.90	1.05E-03	2.98E-02	0.90
Tpbg	chr9:85735986-85740662	5.07	9.48	1.00E-04	4.56E-03	0.90
Kif1a	chr1:94912032-94998442	9.29	17.41	5.00E-05	2.50E-03	0.91
Smoc1	chr12:82120441-82287401	3.28	6.18	1.80E-03	4.51E-02	0.92
Igfbp2	chr1:72871053-72899045	223.16	420.79	5.00E-05	2.50E-03	0.92
Gstm1	chr3:107815167-107820891	12.07	22.79	7.00E-04	2.15E-02	0.92
Sbk1	chr7:133416132-133438513	4.88	9.22	1.50E-04	6.36E-03	0.92
Frzb	chr2:80252126-80287553	6.44	12.21	2.00E-04	7.94E-03	0.92
Fat3	chr9:15714636-16182675	1.07	2.03	1.50E-04	6.36E-03	0.93
Thsd7a	chr6:12261607-12699253	2.27	4.32	5.00E-05	2.50E-03	0.93
Lmo7	chr14:102129144-102333910	3.24	6.19	5.00E-05	2.50E-03	0.93
Crispld2	chr8:122516369-122576693	6.10	11.67	5.00E-05	2.50E-03	0.94
Sfrp1	chr8:24521973-24560104	18.93	36.25	5.00E-05	2.50E-03	0.94
A830080D01Rik	chrX:155970599-156031013	2.15	4.14	1.75E-03	4.41E-02	0.94
Dkk1	chr19:30620373-30623986	7.78	15.00	2.00E-04	7.94E-03	0.95
2510009E07Rik	chr16:21649117-21694738	2.71	5.23	2.50E-04	9.49E-03	0.95
Fzd3	chr14:65811277-65881300	3.16	6.11	5.00E-05	2.50E-03	0.95
Nr2f2	chr7:77496835-77556032	2.90	5.63	7.00E-04	2.15E-02	0.96
Atp10a	chr7:65913571-66084796	1.91	3.78	7.50E-04	2.27E-02	0.98
Syt11	chr3:88548622-88576521	5.33	10.59	5.00E-05	2.50E-03	0.99
Plagl1	chr10:12810593-12851501	54.21	107.83	5.00E-05	2.50E-03	0.99
Dsc2	chr18:20189298-20218006	6.72	13.40	5.00E-05	2.50E-03	1.00
D430019H16Rik	chr12:106692065-106731305	1.72	3.43	9.00E-04	2.63E-02	1.00
Mbnl3	chrX:48466670-48559009	0.92	1.83	1.55E-03	4.03E-02	1.00

Sema3e	chr5:14025275-14256689	2.48	4.99	5.00E-05	2.50E-03	1.01
Arnt2	chr7:91394788-91558469	2.04	4.11	1.50E-04	6.36E-03	1.01
Syt7	chr19:10463579-10527671	1.86	3.78	1.50E-04	6.36E-03	1.02
Sfrp2	chr3:83570242-83578236	6.14	12.49	3.50E-04	1.26E-02	1.03
Col22a1	chr15:71628905-71864657	2.02	4.12	3.00E-04	1.11E-02	1.03
Hbb-y	chr7:111000267-111001721	90.25	185.04	5.00E-05	2.50E-03	1.04
Ptn	chr6:36664884-36761361	12.51	25.83	5.00E-05	2.50E-03	1.05
Bai2	chr4:129662321-129699877	1.40	2.94	1.45E-03	3.82E-02	1.07
Magel2	chr7:69521864-69526526	1.55	3.24	9.50E-04	2.74E-02	1.07
Grem2	chr1:176763915-176851950	3.53	7.44	1.00E-04	4.56E-03	1.07
Prkcz	chr4:154634228-154735500	1.79	3.80	7.00E-04	2.15E-02	1.09
Tpd52	chr3:8929435-9004515	8.47	17.99	5.00E-05	2.50E-03	1.09
Slc4a5	chr6:83187368-83254939	1.67	3.54	4.00E-04	1.40E-02	1.09
Slc8a2	chr7:16715648-16745860	1.96	4.16	5.50E-04	1.78E-02	1.09
Bnc1	chr7:89111547-89137185	3.62	7.76	5.00E-05	2.50E-03	1.10
Trim2	chr3:83964360-84108697	1.22	2.62	3.00E-04	1.11E-02	1.11
Dclk2	chr3:86590071-86724806	2.16	4.67	2.00E-04	7.94E-03	1.11
Mapk8ip1	chr2:92223821-92241420	3.73	8.06	1.00E-04	4.56E-03	1.11
Adamtsl1	chr4:85699818-86074286	1.08	2.33	3.00E-04	1.11E-02	1.11
Epha7	chr4:28740294-28894649	5.19	11.23	5.00E-05	2.50E-03	1.11
Cdon	chr9:35259660-35315237	11.71	25.37	5.00E-05	2.50E-03	1.11
Ltbp4	chr7:28090159-28122631	5.45	11.82	5.00E-05	2.50E-03	1.12
Camsap3	chr8:3587449-3609738	4.31	9.36	5.00E-05	2.50E-03	1.12
Pgap1	chr1:54529843-54614528	2.05	4.46	5.00E-05	2.50E-03	1.12
Lox	chr18:52675724-52689362	4.08	8.89	1.00E-04	4.56E-03	1.12
Prom1	chr5:44384860-44492975	11.51	25.09	5.00E-05	2.50E-03	1.12
Wfdc2	chr2:164388215-164394006	13.61	29.70	9.00E-04	2.63E-02	1.13
Hspa4l	chr3:40549534-40600019	2.80	6.12	5.50E-04	1.78E-02	1.13
Kbtbd11	chr8:15011024-15033332	1.01	2.22	4.00E-04	1.40E-02	1.13
Col9a2	chr4:120712170-120727930	12.56	27.63	5.00E-05	2.50E-03	1.14
Srgap3	chr6:112667965-112897260	2.23	4.91	5.00E-05	2.50E-03	1.14
Mycl1	chr4:122673341-122679723	2.51	5.54	5.00E-05	2.50E-03	1.14
Megf6	chr4:153544821-153649830	3.83	8.47	5.00E-05	2.50E-03	1.14
Zfp534	chr4:147047611-147076662	3.47	7.72	5.00E-05	2.50E-03	1.15
Pim2	chrX:7455431-7460558	4.53	10.09	3.00E-04	1.11E-02	1.16
Col6a3	chr1:92663434-92740548	4.13	9.26	5.00E-05	2.50E-03	1.17
Slc6a15	chr10:102830441-102882013	1.61	3.64	1.45E-03	3.82E-02	1.18
Stag3	chr5:138721736-138753621	3.81	8.60	5.00E-05	2.50E-03	1.18
Sema3d	chr5:12383165-12588943	1.06	2.40	5.50E-04	1.78E-02	1.18
Fmod	chr1:135934091-135944854	2.68	6.13	4.00E-04	1.40E-02	1.19
Mkrn1	chr6:39347819-39370368	9.75	22.42	5.00E-05	2.50E-03	1.20
Sorl1	chr9:41772812-41932372	3.19	7.36	5.00E-05	2.50E-03	1.21
Nptx2	chr5:145306755-145318347	1.88	4.34	1.25E-03	3.41E-02	1.21
Uchl1	chr5:67067359-67078473	6.04	14.02	8.50E-04	2.53E-02	1.21

Sema3a	chr5:13396783-13603485	1.41	3.29	5.00E-05	2.50E-03	1.22
Sdc4	chr2:164249746-164268688	11.55	26.93	5.00E-05	2.50E-03	1.22
Rhpn2	chr7:36119255-36181786	3.14	7.34	1.50E-04	6.36E-03	1.22
Cdhr1	chr14:37891034-37911497	1.13	2.66	7.00E-04	2.15E-02	1.24
Zfp503	chr14:22803183-22808823	4.30	10.16	5.00E-05	2.50E-03	1.24
Gli1	chr10:126760782-126778635	4.81	11.42	5.00E-05	2.50E-03	1.25
Dnm1	chr2:32163990-32208824	1.35	3.20	2.00E-04	7.94E-03	1.25
Wwc1	chr11:35651913-35793591	3.52	8.43	5.00E-05	2.50E-03	1.26
Zbtb16	chr9:48462401-48644050	2.07	4.96	5.00E-05	2.50E-03	1.26
Dnaaf3	chr7:4469909-4484044	3.67	8.86	6.50E-04	2.05E-02	1.27
Ret	chr6:118101765-118147762	0.88	2.14	5.00E-04	1.67E-02	1.28
Slitrk4	chrX:61522618-61530171	0.94	2.29	1.90E-03	4.71E-02	1.28
Nup210	chr6:90963060-91066820	6.56	15.92	5.00E-05	2.50E-03	1.28
Rims3	chr4:120550473-120569173	1.17	2.84	4.50E-04	1.54E-02	1.28
Ccdc88c	chr12:102149752-102267193	1.73	4.22	1.00E-04	4.56E-03	1.29
Epha1	chr6:42308485-42330557	1.18	2.88	1.65E-03	4.22E-02	1.29
Pnpla3	chr15:83998245-84019951	1.09	2.67	7.00E-04	2.15E-02	1.29
Lfng	chr5:141083294-141091499	2.40	5.90	1.50E-04	6.36E-03	1.30
-	chr13:19617298-19620037	2.47	6.06	2.00E-04	7.94E-03	1.30
Dtx4	chr19:12540825-12576486	1.57	3.87	5.00E-05	2.50E-03	1.30
Phf19	chr2:34749274-34769496	1.09	2.68	1.00E-03	2.86E-02	1.30
Plekha7	chr7:123267104-123333355	5.00	12.33	5.00E-05	2.50E-03	1.30
Upk3b	chr5:136514365-136520863	6.82	16.85	5.00E-05	2.50E-03	1.30
Clip4	chr17:72119030-72213550	1.17	2.89	1.50E-03	3.92E-02	1.31
Usp44	chr10:93294299-93320832	1.48	3.67	1.40E-03	3.74E-02	1.31
Igsf9	chr1:174412343-174429008	6.51	16.20	5.00E-05	2.50E-03	1.31
Zfhx4	chr3:5177827-5415855	0.90	2.26	5.00E-05	2.50E-03	1.32
Zswim5	chr4:116550006-116661710	0.80	2.00	9.50E-04	2.74E-02	1.32
Slc4a1	chr11:102210133-102226595	1.60	4.06	1.00E-04	4.56E-03	1.34
Adam11	chr11:102622752-102641576	0.78	1.98	8.00E-04	2.41E-02	1.34
Pdzd4	chrX:71038423-71070308	2.11	5.39	1.00E-04	4.56E-03	1.35
Sept3	chr15:82105364-82124872	2.77	7.13	5.00E-05	2.50E-03	1.36
1700001L05Rik	chr15:83184276-83197727	0.60	1.54	1.05E-03	2.98E-02	1.36
Gldc	chr19:30172930-30249931	8.30	21.36	5.00E-05	2.50E-03	1.36
Fbxl16	chr17:25946029-25958210	0.87	2.26	2.00E-03	4.91E-02	1.37
Phf21b	chr15:84615805-84686559	1.37	3.54	1.40E-03	3.74E-02	1.37
Celsr1	chr15:85729187-85864207	2.19	5.69	5.00E-05	2.50E-03	1.37
Rgma	chr7:80458691-80565871	2.12	5.50	5.00E-05	2.50E-03	1.38
Rab11fip4	chr11:79404713-79511635	2.67	6.95	4.00E-04	1.40E-02	1.38
Cldn6	chr17:23816331-23819414	9.36	24.38	5.00E-05	2.50E-03	1.38
Cadm4	chr7:25267041-25289552	2.35	6.12	4.00E-04	1.40E-02	1.38
Zic2	chr14:122874605-122879550	1.97	5.17	1.50E-04	6.36E-03	1.39
Rfx2	chr17:56915319-56970431	0.92	2.44	1.00E-03	2.86E-02	1.40
Pappa	chr4:64785207-65018543	1.42	3.74	5.00E-05	2.50E-03	1.40

Slc4a8	chr15:100592177-100654402	0.55	1.47	2.50E-04	9.49E-03	1.41
Ldhb	chr6:142438768-142456463	30.96	82.29	5.00E-05	2.50E-03	1.41
4930470H14Rik	chr17:4044657-4082995	7.45	19.94	5.00E-05	2.50E-03	1.42
Myh14	chr7:51861172-51926213	0.49	1.32	7.50E-04	2.27E-02	1.42
Rragb	chrX:149574500-149606486	0.72	1.95	6.00E-04	1.93E-02	1.43
Spint2	chr7:30041348-30066996	7.60	20.60	5.00E-05	2.50E-03	1.44
Col2a1	chr15:97806032-97835155	21.41	58.08	5.00E-05	2.50E-03	1.44
Dtna	chr18:23573915-23818215	1.18	3.21	3.50E-04	1.26E-02	1.44
Nrk	chrX:135448968-135545068	4.28	11.73	5.00E-05	2.50E-03	1.46
Glb1l2	chr9:26570628-26614002	1.01	2.78	6.00E-04	1.93E-02	1.46
Cyp2s1	chr7:26587494-26601549	1.49	4.13	5.00E-04	1.67E-02	1.47
Map7	chr10:19868725-20001396	1.59	4.47	5.00E-05	2.50E-03	1.49
Cbx7	chr15:79746236-79763076	3.52	9.92	5.00E-05	2.50E-03	1.50
Robo3	chr9:37223629-37240760	0.97	2.74	7.00E-04	2.15E-02	1.50
Pcsk9	chr4:106114938-106136930	1.27	3.59	4.00E-04	1.40E-02	1.50
Ephb1	chr9:101824457-102257023	0.86	2.44	2.00E-04	7.94E-03	1.50
Cbx7	chr15:79746236-79763076	1.50	4.27	9.50E-04	2.74E-02	1.51
Elavl2	chr4:90917456-91066675	1.01	2.90	3.00E-04	1.11E-02	1.51
Scube1	chr15:83433012-83555469	6.87	19.73	5.00E-05	2.50E-03	1.52
Nlrp1a	chr11:70904698-70958206	1.24	3.58	5.00E-05	2.50E-03	1.54
Lum	chr10:97028134-97035337	30.29	88.17	5.00E-05	2.50E-03	1.54
F2rl1	chr13:96281683-96295195	2.47	7.21	5.00E-05	2.50E-03	1.55
Cmah	chr13:24419288-24569154	2.03	5.95	5.00E-05	2.50E-03	1.55
Sfmbt2	chr2:10292077-10516880	1.76	5.16	5.00E-05	2.50E-03	1.55
Adamts19	chr18:58996417-59213332	0.70	2.05	5.00E-04	1.67E-02	1.55
Vtn	chr11:78312621-78315827	2.51	7.36	4.50E-04	1.54E-02	1.55
Cgn	chr3:94563991-94590437	2.37	7.03	5.00E-05	2.50E-03	1.57
Epb4.1l4a	chr18:33955980-34166860	1.74	5.17	5.00E-05	2.50E-03	1.57
Celsr2	chr3:108193765-108218412	1.98	5.90	5.00E-05	2.50E-03	1.58
Map3k9	chr12:82815936-82882157	0.45	1.33	4.50E-04	1.54E-02	1.58
Cobl	chr11:12136678-12364963	0.45	1.35	1.75E-03	4.41E-02	1.58
Fzd5	chr1:64777131-64784324	1.88	5.64	5.00E-05	2.50E-03	1.58
Dppa4	chr16:48283847-48294405	3.61	10.85	3.00E-04	1.11E-02	1.59
Frem2	chr3:53317859-53461277	5.29	16.01	5.00E-05	2.50E-03	1.60
Mt1	chr8:96702988-96704227	22.15	67.24	5.00E-05	2.50E-03	1.60
Rab11fip4	chr11:79404713-79511635	3.11	9.50	5.00E-05	2.50E-03	1.61
Kif21a	chr15:90763706-90880379	2.92	8.92	5.00E-05	2.50E-03	1.61
Lrp2	chr2:69262391-69424124	6.40	19.65	5.00E-05	2.50E-03	1.62
Slc45a3	chr1:133859491-133879549	0.48	1.47	1.55E-03	4.03E-02	1.63
Camkv	chr9:107838250-107852022	1.56	4.88	2.00E-04	7.94E-03	1.65
Dnahc8	chr17:30763880-31012209	0.43	1.34	5.00E-05	2.50E-03	1.65
Cpz	chr5:35844866-35868275	3.14	9.95	5.00E-05	2.50E-03	1.66
Lrrn1	chr6:107479719-107520222	0.68	2.16	2.50E-04	9.49E-03	1.67
Pcsk6	chr7:73007021-73195272	1.12	3.57	5.00E-05	2.50E-03	1.67

Dcn	chr10:96942133-96980796	6.02	19.18	5.00E-05	2.50E-03	1.67
Nuak2	chr1:134212701-134230065	1.44	4.60	5.00E-05	2.50E-03	1.67
Clstn3	chr6:124380773-124414802	1.46	4.67	5.00E-05	2.50E-03	1.67
Galnt16	chr12:81619976-81704883	1.23	3.97	5.00E-05	2.50E-03	1.68
Fcho1	chr8:74232285-74249580	1.13	3.64	5.00E-05	2.50E-03	1.69
Crabp2	chr3:87750395-87757294	14.17	46.02	5.00E-05	2.50E-03	1.70
Slc35f2	chr9:53619341-53665968	3.66	12.02	5.00E-05	2.50E-03	1.72
Pramef12	chr4:143981576-143998367	0.84	2.77	6.50E-04	2.05E-02	1.72
Rspo2	chr15:42852340-43002364	0.70	2.30	1.60E-03	4.12E-02	1.72
AI414108	chr9:27160269-27165128	1.32	4.34	5.00E-05	2.50E-03	1.72
Hook1	chr4:95626401-95692055	3.03	10.23	5.00E-05	2.50E-03	1.75
Dll1	chr17:15504317-15512787	1.08	3.65	5.00E-05	2.50E-03	1.75
Alpl	chr4:137297646-137352292	8.15	27.70	5.00E-05	2.50E-03	1.76
Gm13051	chr4:146064712-146094024	1.72	5.86	5.00E-05	2.50E-03	1.77
Fabp3	chr4:129986021-129992707	13.10	44.92	5.00E-05	2.50E-03	1.78
St14	chr9:30896174-30939384	1.84	6.32	5.00E-05	2.50E-03	1.78
Trh	chr6:92192055-92194642	11.81	40.60	5.00E-05	2.50E-03	1.78
-	chr9:75318079-75322594	0.50	1.72	5.00E-04	1.67E-02	1.78
Spint1	chr2:119063095-119075249	2.79	9.63	5.00E-05	2.50E-03	1.79
Mylpf	chr7:134355121-134357801	3.51	12.13	4.50E-04	1.54E-02	1.79
Adam23	chr1:63492477-63643089	0.71	2.45	5.00E-05	2.50E-03	1.80
Slc35f1	chr10:52410306-52831428	0.52	1.82	2.50E-04	9.49E-03	1.80
Cenpv	chr11:62338445-62352763	0.72	2.52	5.00E-05	2.50E-03	1.80
-	chr9:27159455-27160230	2.20	7.72	1.60E-03	4.12E-02	1.81
Chchd10	chr10:75398317-75400479	3.28	11.49	4.00E-04	1.40E-02	1.81
Spock2	chr10:59569004-59597899	1.23	4.31	5.00E-05	2.50E-03	1.81
Slit1	chr19:41674746-41818346	2.13	7.52	5.00E-05	2.50E-03	1.82
Slc4a11	chr2:130509843-130523255	0.51	1.82	3.50E-04	1.26E-02	1.83
Pou3f1	chr4:124334888-124337899	0.40	1.43	1.25E-03	3.41E-02	1.83
Cdh6	chr15:12958488-13103394	3.31	11.83	5.00E-05	2.50E-03	1.84
Cdh6	chr15:12958488-13103394	2.02	7.23	5.00E-05	2.50E-03	1.84
Syt9	chr7:114514303-114692169	0.47	1.68	3.00E-04	1.11E-02	1.84
Atp1a3	chr7:25763187-25790914	0.48	1.73	7.00E-04	2.15E-02	1.86
Glyctk	chr9:106055190-106062000	0.41	1.50	5.00E-04	1.67E-02	1.86
Grik3	chr4:125168074-125391417	0.68	2.47	5.00E-05	2.50E-03	1.87
Nr5a2	chr1:138740160-138857025	0.40	1.46	1.30E-03	3.52E-02	1.88
Grin1	chr2:25145430-25174683	0.47	1.72	3.50E-04	1.26E-02	1.88
Foxa2	chr2:147868613-147872705	0.96	3.54	6.50E-04	2.05E-02	1.88
Tox3	chr8:92771010-92872151	1.09	4.02	5.00E-05	2.50E-03	1.88
Fgfbp3	chr19:36992039-36994089	3.30	12.20	5.00E-05	2.50E-03	1.89
Pax6	chr2:105376236-105540183	0.86	3.19	5.00E-05	2.50E-03	1.89
-	chr19:41672351-41674323	1.26	4.70	4.00E-04	1.40E-02	1.90
Wnk2	chr13:49131670-49243383	0.51	1.92	5.00E-05	2.50E-03	1.91
Shh	chr5:28783379-29045749	3.16	12.03	5.00E-05	2.50E-03	1.93

-	chr9:27153357-27159388	0.81	3.09	5.00E-05	2.50E-03	1.93
Ybx2	chr11:69749400-69755101	2.24	8.56	5.00E-05	2.50E-03	1.93
Plch1	chr3:63500155-63654913	0.81	3.10	5.00E-05	2.50E-03	1.94
Abcc4	chr14:118881913-119105441	0.96	3.74	5.00E-05	2.50E-03	1.96
Fst	chr13:115242469-115248938	2.10	8.17	5.00E-05	2.50E-03	1.96
Ttc39b	chr4:82866204-82970093	0.74	2.89	5.00E-05	2.50E-03	1.96
Sema5b	chr16:35541447-35664344	0.88	3.44	5.00E-05	2.50E-03	1.96
Jam2	chr16:84774367-84826620	0.65	2.61	1.50E-04	6.36E-03	2.00
Zic3	chrX:55283804-55289807	1.53	6.18	5.00E-05	2.50E-03	2.01
Islr2	chr9:58044103-58056615	0.46	1.86	1.00E-04	4.56E-03	2.01
Sox21	chr14:118632455-118636252	0.43	1.77	1.50E-04	6.36E-03	2.03
Eya2	chr2:165420527-165597227	0.79	3.23	5.00E-05	2.50E-03	2.03
Igfbpl1	chr4:45822378-45839699	0.68	2.82	1.50E-04	6.36E-03	2.04
Slain1	chr14:104049459-104104016	0.41	1.69	7.00E-04	2.15E-02	2.05
Mpzl2	chr9:44850426-44862126	0.40	1.65	7.50E-04	2.27E-02	2.06
Elovl2	chr13:41277750-41315772	0.71	2.96	5.00E-05	2.50E-03	2.06
Grhl2	chr15:37162790-37293323	0.39	1.65	1.00E-04	4.56E-03	2.06
Kcnk1	chr8:128519001-128554585	1.30	5.47	5.00E-05	2.50E-03	2.07
Ephx2	chr14:66703208-66743359	1.23	5.20	1.00E-04	4.56E-03	2.08
Lefty1	chr1:182865169-182868532	1.09	4.65	2.50E-04	9.49E-03	2.09
Cdh1	chr8:109127267-109194146	7.05	30.06	5.00E-05	2.50E-03	2.09
Mreg	chr1:72205806-72258881	1.53	6.53	5.00E-05	2.50E-03	2.10
Lmx1a	chr1:169619688-169778864	0.94	4.02	5.00E-05	2.50E-03	2.10
Map2	chr1:66221902-66489157	0.80	3.44	5.00E-05	2.50E-03	2.10
Otx2	chr14:49276684-49282547	5.19	22.75	5.00E-05	2.50E-03	2.13
Tdgf1	chr9:110842111-110848662	3.34	14.76	5.00E-05	2.50E-03	2.14
Epcam	chr17:88035318-88050467	4.43	19.85	5.00E-05	2.50E-03	2.16
Sox2	chr3:34459302-34576915	7.80	35.12	5.00E-05	2.50E-03	2.17
Hap1	chr11:100208640-100217455	1.35	6.12	5.00E-05	2.50E-03	2.18
Sox3	chrX:58144540-58146605	0.42	1.88	5.50E-04	1.78E-02	2.18
Gyltl1b	chr2:92205202-92211193	1.76	8.01	5.00E-05	2.50E-03	2.19
Esrp1	chr4:11259184-11313930	0.94	4.26	5.00E-05	2.50E-03	2.19
Pax3	chr1:78097841-78193711	0.30	1.36	2.50E-04	9.49E-03	2.20
H2-Bl	chr17:36217133-36221194	0.85	3.95	1.10E-03	3.09E-02	2.21
Fez1	chr9:36651243-36686225	0.83	3.87	9.00E-04	2.63E-02	2.22
BC068157	chr8:4209542-4217312	0.61	2.88	5.00E-05	2.50E-03	2.23
Irs4	chrX:138145540-138159760	0.43	2.03	5.00E-05	2.50E-03	2.25
Lhx2	chr2:38206827-38225248	0.93	4.51	5.00E-05	2.50E-03	2.28
Cldn4	chr5:135420992-135422804	3.95	19.31	5.00E-05	2.50E-03	2.29
Mapk13	chr17:28906261-28915649	0.90	4.42	2.00E-04	7.94E-03	2.29
Kif5a	chr10:126662750-126700419	0.51	2.49	5.00E-05	2.50E-03	2.29
Calca	chr7:121774991-121779871	0.82	4.03	1.35E-03	3.63E-02	2.29
B4galnt3	chr6:120150837-120244577	0.58	2.87	5.00E-04	1.67E-02	2.31
Cbs	chr17:31749567-31774150	0.52	2.58	5.00E-05	2.50E-03	2.32

C1s	chr6:124480361-124492377	0.28	1.42	5.50E-04	1.78E-02	2.33
-	chr8:3585435-3587396	0.47	2.37	6.00E-04	1.93E-02	2.34
Mgat5b	chr11:116780176-116848258	0.29	1.48	1.50E-04	6.36E-03	2.34
Enpp3	chr10:24493619-24556001	2.39	12.15	5.00E-05	2.50E-03	2.35
Aqp3	chr4:41039756-41045216	0.86	4.42	5.00E-04	1.67E-02	2.35
Cldn7	chr11:69778280-69781388	1.77	9.05	5.00E-05	2.50E-03	2.36
Zfp296	chr7:20162635-20166005	0.45	2.29	1.80E-03	4.51E-02	2.36
Tfap2a	chr13:40811043-40829192	0.62	3.18	5.00E-05	2.50E-03	2.36
Rnf17	chr14:57021533-57143868	0.34	1.78	5.00E-05	2.50E-03	2.38
Ttyh1	chr7:4025726-4088528	0.31	1.62	5.00E-05	2.50E-03	2.40
Dmrt1	chr19:25580195-25678818	0.42	2.26	3.50E-04	1.26E-02	2.44
L1td1	chr4:98393444-98405177	1.34	7.26	5.00E-05	2.50E-03	2.44
-	chr14:49282574-49287385	1.05	5.70	5.00E-05	2.50E-03	2.44
Pipox	chr11:77694116-77707374	1.02	5.53	5.00E-05	2.50E-03	2.44
Cpne5	chr17:29293465-29374735	0.29	1.61	5.00E-05	2.50E-03	2.46
Srcin1	chr11:97370653-97436440	0.33	1.81	5.00E-05	2.50E-03	2.47
Erbb3	chr10:128004594-128026557	1.11	6.16	3.00E-04	1.11E-02	2.47
Ntn1	chr11:68022865-68200328	3.13	17.49	5.00E-05	2.50E-03	2.48
Fgfbp1	chr5:44370096-44373038	0.81	4.61	1.50E-04	6.36E-03	2.50
B4galnt3	chr6:120150837-120244577	0.50	2.87	6.50E-04	2.05E-02	2.52
C130021I20Rik	chr2:33496712-33501823	0.64	3.68	5.00E-05	2.50E-03	2.53
Nell2	chr15:95049880-95359137	0.26	1.50	3.00E-04	1.11E-02	2.55
Kcnj3	chr2:55289566-55450492	0.38	2.31	1.20E-03	3.30E-02	2.60
Cldn3	chr5:135462083-135477220	0.59	3.61	9.50E-04	2.74E-02	2.60
Fmr1nb	chrX:66015013-66057735	0.75	4.57	1.00E-03	2.86E-02	2.61
Mir135a-2	chr10:91534360-91534930	5.20	32.18	5.00E-05	2.50E-03	2.63
Liph	chr16:21953890-21995615	0.34	2.11	2.00E-04	7.94E-03	2.64
Erbb3	chr10:128004594-128026557	0.84	5.22	5.00E-05	2.50E-03	2.64
Fam181b	chr7:100228388-100230231	0.71	4.46	1.00E-04	4.56E-03	2.66
Gpr98	chr13:81234066-81772143	0.33	2.11	5.00E-05	2.50E-03	2.66
Six3	chr17:86020173-86025531	0.56	3.65	5.00E-05	2.50E-03	2.70
Six3os1	chr17:86001271-86017736	0.45	3.03	6.50E-04	2.05E-02	2.74
Tfcp2l1	chr1:120524521-120581745	0.45	3.00	5.00E-05	2.50E-03	2.75
Otx1	chr11:21894766-21901654	0.34	2.29	1.50E-04	6.36E-03	2.76
Tcea3	chr4:135803871-135830814	0.73	4.96	8.00E-04	2.41E-02	2.76
Gjb3	chr4:127002478-127008080	1.58	10.71	5.00E-05	2.50E-03	2.76
Morc1	chr16:48431349-48631018	0.26	1.80	3.50E-04	1.26E-02	2.77
Mt2	chr8:96696517-96697467	11.26	77.75	5.00E-05	2.50E-03	2.79
Ildr2	chr1:168184269-168246963	0.95	6.65	5.00E-05	2.50E-03	2.80
Ap1m2	chr9:21099900-21116777	0.34	2.40	1.00E-03	2.86E-02	2.80
4930500J02Rik	chr2:104399333-104411586	1.92	13.64	9.00E-04	2.63E-02	2.83
Gstp2	chr19:4040287-4042221	1.62	11.68	1.30E-03	3.52E-02	2.85
Gpat2	chr2:127250934-127261949	0.40	2.88	1.50E-04	6.36E-03	2.85
B3gnt7	chr1:88199795-88203880	1.39	10.04	5.00E-05	2.50E-03	2.86

Elf3	chr1:137150150-137155049	0.66	4.83	5.00E-05	2.50E-03	2.87
Six3os1	chr17:86001271-86017736	0.99	7.33	5.00E-05	2.50E-03	2.88
Nrcam	chr12:45429871-45702833	0.66	4.91	5.00E-05	2.50E-03	2.89
Dbx1	chr7:56886868-56892205	0.51	3.88	3.00E-04	1.11E-02	2.92
D7Ertd143e	chr7:3217861-3221016	0.61	4.87	5.00E-05	2.50E-03	3.01
Mpped1	chr15:83610452-83688904	0.21	1.72	5.00E-05	2.50E-03	3.01
Pcdh8	chr14:80166578-80171119	0.64	5.20	5.00E-05	2.50E-03	3.02
Sim2	chr16:94085504-94348638	0.47	3.80	5.00E-05	2.50E-03	3.02
Miat	chr5:112642247-112657968	1.81	14.87	5.00E-05	2.50E-03	3.04
Nanog	chr6:122657585-122664639	2.20	18.08	5.00E-05	2.50E-03	3.04
Mlxipl	chr5:135582760-135614252	0.18	1.50	3.50E-04	1.26E-02	3.05
Foxb1	chr9:69605516-69608747	0.19	1.63	1.70E-03	4.33E-02	3.07
Dppa5a	chr9:78214860-78216006	35.20	298.69	5.00E-05	2.50E-03	3.09
Wnt7b	chr15:85365866-85424138	0.26	2.20	5.00E-05	2.50E-03	3.09
2410141K09Rik	chr13:66519049-66542054	2.10	18.01	1.15E-03	3.19E-02	3.10
Wnt1	chr15:98620287-98624261	0.64	5.51	5.00E-05	2.50E-03	3.10
Fezf2	chr14:13174405-13179290	0.43	3.69	2.00E-04	7.94E-03	3.12
Gm13247	chr4:145651165-145696039	0.53	4.69	5.00E-05	2.50E-03	3.14
Nkx2-1	chr12:57632923-57637895	0.36	3.25	2.00E-04	7.94E-03	3.16
Rfx4	chr10:84218792-84369283	0.33	2.99	5.00E-05	2.50E-03	3.19
Ap3b2	chr7:88605284-88638811	0.29	2.64	5.00E-05	2.50E-03	3.20
Smc1b	chr15:84895118-84962387	0.23	2.12	1.00E-04	4.56E-03	3.21
Wnt8b	chr19:44567961-44590041	0.30	2.83	1.00E-04	4.56E-03	3.22
Ano9	chr7:148287117-148303705	0.31	2.86	1.00E-04	4.56E-03	3.22
Sox1	chr8:12385770-12436732	0.34	3.21	5.00E-05	2.50E-03	3.24
Tfap2c	chr2:172375092-172384121	0.41	3.92	5.00E-05	2.50E-03	3.25
Gm13242	chr4:145126547-145419626	0.55	5.21	1.60E-03	4.12E-02	3.26
Zfp42	chr8:44380420-44392363	1.04	10.10	5.00E-05	2.50E-03	3.28
Esrrb	chr12:87702066-87862578	0.65	6.28	5.00E-05	2.50E-03	3.28
Mcf2	chrX:57309132-57400820	0.16	1.51	1.55E-03	4.03E-02	3.28
Trap1a	chrX:135774764-135892277	1.38	13.49	5.00E-05	2.50E-03	3.29
Aire	chr10:77492766-77526360	0.31	3.04	5.00E-05	2.50E-03	3.29
Slc28a1	chr7:88259684-88315302	0.23	2.23	2.00E-04	7.94E-03	3.29
Esrp1	chr4:11259184-11313930	0.23	2.27	1.40E-03	3.74E-02	3.31
Pou5f1	chr17:35642976-35647722	3.87	39.95	5.00E-05	2.50E-03	3.37
Folr1	chr7:109006844-109019302	0.43	4.59	5.00E-05	2.50E-03	3.41
Gpa33	chr1:168060590-168096641	0.30	3.20	4.00E-04	1.40E-02	3.41
Sptbn2	chr19:4711222-4752352	0.39	4.13	5.00E-05	2.50E-03	3.41
Utf1	chr7:147129754-147131011	0.85	9.21	2.00E-04	7.94E-03	3.43
Alox15	chr11:70157648-70165533	0.29	3.16	1.00E-04	4.56E-03	3.45
-	chr9:118308486-118313594	0.15	1.77	2.50E-04	9.49E-03	3.56
Fgf4	chr7:152047290-152051148	0.36	4.35	5.00E-05	2.50E-03	3.61
BC024139, Eppk1	chr15:75931917-75956986	0.12	1.47	2.50E-04	9.49E-03	3.62
Tdrd12	chr7:36278628-36322763	0.40	5.02	1.60E-03	4.12E-02	3.65

AU018091	chr7:3154659-3169204	0.88	11.07	5.00E-05	2.50E-03	3.65
Gm2381	chr7:50067562-50122604	0.19	2.64	5.00E-04	1.67E-02	3.77
Gdf3	chr6:122555420-122560089	0.28	4.04	5.00E-04	1.67E-02	3.85
Tdh	chr14:64111183-64127929	0.75	10.95	5.00E-05	2.50E-03	3.86
-	chr13:98278330-98283216	0.11	1.78	6.50E-04	2.05E-02	4.05
4930500J02Rik	chr2:104399333-104411586	0.10	1.60	9.50E-04	2.74E-02	4.06
-	chr13:98252715-98274765	0.12	2.13	5.00E-05	2.50E-03	4.11
Gm10324	chr13:66214388-66223772	0.22	6.41	9.50E-04	2.74E-02	4.87
-	chr15:96991387-96991826	0.00	2.55	5.00E-05	2.50E-03	Infinite
-	chr7:36255279-36256279	0.00	38.35	2.00E-03	4.91E-02	Infinite

A.5 GO Term Analysis Results Tables

Table A.5: Significantly overrepresented, pruned, biological process GO terms from the differential expression analysis. Terms are sorted by ascending adjusted p-value and then by descending Z-score.

Ontology Name (Ontology-ID)	Percent Changed	Z-Score	Adjusted p-value
developmental process(GO:0032502)	8.06	16.86	1.01E-02
cell adhesion(GO:0007155)	13.56	14.43	1.01E-02
regulation of developmental process(GO:0050793)	9.63	12.97	1.01E-02
regulation of cell migration(GO:0030334)	15.00	12.17	1.01E-02
regulation of response to stimulus(GO:0048583)	7.94	11.07	1.01E-02
positive regulation of biological process(GO:0048518)	6.71	11.05	1.01E-02
regulation of cell proliferation(GO:0042127)	9.29	10.83	1.01E-02
response to external stimulus(GO:0009605)	9.87	10.55	1.01E-02
negative regulation of cellular process(GO:0048523)	6.88	10.17	1.01E-02
negative regulation of locomotion(GO:0040013)	20.35	10.17	1.01E-02
locomotion(GO:0040011)	10.32	10.01	1.01E-02
regulation of transcription from RNA polymerase II promoter(GO:0006357)	8.62	9.88	1.01E-02
regulation of signaling(GO:0023051)	7.58	9.76	1.01E-02
regulation of cell adhesion(GO:0030155)	14.16	9.33	1.01E-02
tissue remodeling(GO:0048771)	22.39	8.75	1.01E-02
regulation of gamma-delta T cell activation(GO:0046643)	66.67	8.68	1.01E-02
regulation of cellular component organization(GO:0051128)	7.90	8.52	1.01E-02
regulation of cell communication(GO:0010646)	7.58	8.50	1.01E-02
JAK-STAT cascade involved in growth hormone signaling pathway(GO:0060397)	75.00	8.02	1.01E-02
enzyme linked receptor protein signaling pathway(GO:0007167)	10.86	7.98	1.01E-02
cell communication(GO:0007154)	9.53	7.67	1.01E-02
response to fluid shear stress(GO:0034405)	37.50	7.66	1.01E-02
regulation of biological quality(GO:0065008)	6.55	7.56	1.01E-02
cellular response to chemical stimulus(GO:0070887)	7.52	7.32	1.01E-02
circulatory system process(GO:0003013)	16.13	6.94	1.01E-02
immune system process(GO:0002376)	7.70	6.89	1.01E-02
negative regulation of ossification(GO:0030279)	25.93	6.58	1.01E-02
molting cycle process(GO:0022404)	17.14	6.49	1.01E-02
regulation of coagulation(GO:0050818)	19.23	6.43	1.01E-02
response to wounding(GO:0009611)	8.96	6.41	1.01E-02
Wnt receptor signaling pathway involved in somitogenesis(GO:0090244)	50.00	6.40	1.01E-02
response to endogenous stimulus(GO:0009719)	7.29	6.31	1.01E-02
peptidyl-tyrosine modification(GO:0018212)	17.46	6.30	1.01E-02
regulation of kinase activity(GO:0043549)	8.18	6.03	1.01E-02
extracellular structure organization(GO:0043062)	10.94	5.95	1.01E-02
positive regulation of molecular function(GO:0044093)	6.72	5.90	1.01E-02
cell proliferation involved in metanephros development(GO:0072203)	42.86	5.86	1.01E-02
peptide cross-linking via chondroitin 4-sulfate glycosaminoglycan(GO:0019800)	42.86	5.86	1.01E-02
regulation of apoptosis(GO:0042981)	6.46	5.81	1.01E-02

canonical Wnt receptor signaling pathway(GO:0060070)	14.29	5.64	1.01E-02
regulation of cell division(GO:0051302)	15.07	5.63	1.01E-02
signaling(GO:0023052)	9.26	5.51	1.01E-02
regulation of bone remodeling(GO:0046850)	22.22	5.50	1.01E-02
regulation of vasodilation(GO:0042312)	21.43	5.37	1.01E-02
negative regulation of catalytic activity(GO:0043086)	7.36	5.23	1.01E-02
membrane invagination(GO:0010324)	9.28	5.18	1.01E-02
response to hypoxia(GO:0001666)	9.29	5.07	1.01E-02
locomotory behavior(GO:0007626)	10.67	5.06	1.01E-02
cell projection organization(GO:0030030)	7.48	4.88	1.01E-02
regulation of cyclic nucleotide metabolic process(GO:0030799)	10.69	4.74	1.01E-02
female pregnancy(GO:0007565)	14.81	4.74	1.01E-02
regulation of cellular component biogenesis(GO:0044087)	8.30	4.58	1.01E-02
regulation of nucleotide biosynthetic process(GO:0030808)	10.48	4.49	1.01E-02
regulation of muscle system process(GO:0090257)	11.46	4.48	1.01E-02
establishment or maintenance of cell polarity(GO:0007163)	12.68	4.43	1.01E-02
response to pain(GO:0048265)	18.52	4.43	1.01E-02
regulation of actin filament-based process(GO:0032970)	9.86	4.39	1.01E-02
non-canonical Wnt receptor signaling pathway(GO:0035567)	17.24	4.20	1.01E-02
second-messenger-mediated signaling(GO:0019932)	8.59	4.18	1.01E-02
actin filament-based process(GO:0030029)	7.81	4.16	1.01E-02
regulation of binding(GO:0051098)	9.04	4.15	1.01E-02
regulation of calcium ion transport(GO:0051924)	9.63	4.13	1.01E-02
response to growth factor stimulus(GO:0070848)	10.00	4.12	1.01E-02
ossification(GO:0001503)	9.71	3.65	1.01E-02
response to molecule of bacterial origin(GO:0002237)	7.49	3.55	1.01E-02
secretion(GO:0046903)	6.77	3.53	1.01E-02
response to ethanol(GO:0045471)	9.09	3.41	1.01E-02
polyol transport(GO:0015791)	37.50	5.41	1.83E-02
regulation of calcium ion-dependent exocytosis(GO:0017158)	20.00	4.67	1.83E-02
regulation of leukocyte mediated cytotoxicity(GO:0001910)	14.29	3.99	1.83E-02
Ras protein signal transduction(GO:0007265)	10.47	3.72	1.83E-02
regulation of leukocyte mediated immunity(GO:0002703)	9.90	3.72	1.83E-02
proteoglycan metabolic process(GO:0006029)	20.00	4.67	2.56E-02
fluid transport(GO:0042044)	20.00	4.18	2.56E-02
vesicle fusion(GO:0006906)	20.00	4.18	2.56E-02
coagulation(GO:0050817)	11.67	3.63	2.56E-02
protein processing(GO:0016485)	10.53	3.53	2.56E-02
transcription from RNA polymerase II promoter(GO:0006366)	8.33	3.24	2.56E-02
luteinization(GO:0001553)	37.50	5.41	3.26E-02
serotonin transport(GO:0006837)	33.33	5.04	3.26E-02
sexual reproduction(GO:0019953)	30.00	4.72	3.26E-02
MAPKKK cascade(GO:0000165)	8.91	3.16	3.26E-02
regulation of establishment or maintenance of cell polarity(GO:0032878)	37.50	5.41	3.90E-02

succinate metabolic process(GO:0006105)	30.00	4.72	3.90E-02
nitric oxide mediated signal transduction(GO:0007263)	27.27	4.45	3.90E-02
response to X-ray(GO:0010165)	16.00	3.55	3.90E-02
lipid localization(GO:0010876)	21.05	4.33	4.50E-02
peptidyl-tyrosine dephosphorylation(GO:0035335)	12.50	3.57	4.50E-02

Table A.6: Significantly overrepresented, pruned, cellular component GO terms from the differential expression analysis. Terms are sorted by ascending adjusted p-value and then by descending Z-score.

Ontology Name(Ontology-ID)	Percent Changed	Z-Score	Adjusted p-value
extracellular region part(GO:0044421)	9.31	11.34	1.01E-02
plasma membrane part(GO:0044459)	7.58	11.03	1.01E-02
cell surface(GO:0009986)	13.41	10.80	1.01E-02
plasma membrane(GO:0005886)	6.61	10.59	1.01E-02
neuron projection(GO:0043005)	9.38	8.04	1.01E-02
dendrite terminus(GO:0044292)	75.00	8.02	1.01E-02
extracellular region(GO:0005576)	6.71	7.74	1.01E-02
membrane raft(GO:0045121)	10.14	5.54	1.01E-02
cell periphery(GO:0071944)	21.74	4.95	1.01E-02
extrinsic to membrane(GO:0019898)	10.92	4.67	1.01E-02
transport vesicle(GO:0030133)	14.29	4.60	1.01E-02
secretory granule membrane(GO:0030667)	15.56	4.60	1.01E-02
apical part of cell(GO:0045177)	12.05	4.47	1.01E-02
axon part(GO:0033267)	9.68	4.46	1.01E-02
cell fraction(GO:0000267)	5.48	4.30	1.01E-02
synapse(GO:0045202)	7.28	4.25	1.01E-02
synapse part(GO:0044456)	6.95	3.99	1.01E-02
perinuclear region of cytoplasm(GO:0048471)	6.68	3.91	1.01E-02
uropod(GO:0001931)	42.86	5.86	1.83E-02
transcription factor complex(GO:0005667)	6.33	3.03	1.83E-02
cell cortex part(GO:0044448)	9.52	3.20	3.26E-02
cell body(GO:0044297)	6.27	3.06	3.90E-02

Table A.7: Significantly overrepresented, pruned, molecular function GO terms from the differential expression analysis. Terms are sorted by ascending adjusted p-value and then by descending Z-score.

Ontology Name(Ontology-ID)	Percent Changed	Z-Score	Adjusted p-value
protein binding(GO:0005515)	4.93	11.59	1.01E-02
pattern binding(GO:0001871)	18.07	10.69	1.01E-02
calcium ion binding(GO:0005509)	10.87	10.52	1.01E-02
extracellular matrix binding(GO:0050840)	28.95	8.85	1.01E-02
RNA polymerase II regulatory region sequence-specific DNA binding(GO:0000977)	22.22	8.41	1.01E-02
Wnt receptor activity(GO:0042813)	35.00	7.93	1.01E-02
chemorepellent activity(GO:0045499)	60.00	7.09	1.01E-02
retinoic acid binding(GO:0001972)	44.44	6.91	1.01E-02
transmembrane receptor protein kinase activity(GO:0019199)	16.67	6.87	1.01E-02
icosanoid binding(GO:0050542)	50.00	6.40	1.01E-02
sequence-specific DNA binding transcription factor activity(GO:0003700)	7.20	6.38	1.01E-02
inorganic anion exchanger activity(GO:0005452)	33.33	5.82	1.01E-02
core promoter sequence-specific DNA binding(GO:0001046)	23.08	5.64	1.01E-02
chromatin binding(GO:0003682)	9.16	5.23	1.01E-02
enhancer sequence-specific DNA binding(GO:0001158)	23.81	5.26	1.83E-02
scavenger receptor activity(GO:0005044)	14.29	3.99	1.83E-02
miRNA binding(GO:0035198)	42.86	5.86	2.56E-02
receptor signaling protein activity(GO:0005057)	10.26	3.44	2.56E-02
protein tyrosine phosphatase activity(GO:0004725)	9.28	3.30	2.56E-02
water channel activity(GO:0015250)	27.27	4.45	3.26E-02
protein serine/threonine kinase activity(GO:0004674)	5.54	2.95	3.90E-02
GTPase regulator activity(GO:0030695)	5.76	2.83	3.90E-02
axon guidance receptor activity(GO:0008046)	37.50	5.41	4.50E-02

Table A.8: Significantly overrepresented, pruned, biological process GO terms from the genes upregulated in the endocardium. Terms are sorted by ascending adjusted p-value and then by descending Z-score.

Ontology Name (Ontology-ID)	Percent Changed	Z-Score	Adjusted p-value
regulation of erythrocyte differentiation(GO:0045646)	32.26	14.55	1.20E-02
regulation of mast cell differentiation(GO:0060375)	100.00	14.46	1.20E-02
endothelial cell differentiation(GO:0045446)	55.56	13.75	1.20E-02
regulation of gamma-delta T cell activation(GO:0046643)	66.67	13.53	1.20E-02
angiogenesis(GO:0001525)	12.90	13.33	1.20E-02
JAK-STAT cascade involved in growth hormone signaling pathway(GO:0060397)	75.00	12.46	1.20E-02
regulation of cell motility(GO:2000145)	9.07	12.28	1.20E-02
response to fluid shear stress(GO:0034405)	37.50	12.22	1.20E-02
regulation of angiogenesis(GO:0045765)	13.64	11.93	1.20E-02
hemopoiesis(GO:0030097)	15.79	11.89	1.20E-02
negative regulation of protein autophosphorylation(GO:0031953)	50.00	11.64	1.20E-02
regulation of response to stimulus(GO:0048583)	4.57	11.40	1.20E-02
granulocyte differentiation(GO:0030851)	38.46	11.31	1.20E-02
platelet formation(GO:0030220)	44.44	10.93	1.20E-02
vasculogenesis(GO:0001570)	16.44	10.89	1.20E-02
negative regulation of endothelial cell proliferation(GO:0001937)	28.57	10.54	1.20E-02
positive regulation of myeloid cell differentiation(GO:0045639)	16.95	10.12	1.20E-02
positive regulation of cellular component movement(GO:0051272)	9.01	9.63	1.20E-02
tissue remodeling(GO:0048771)	14.93	9.38	1.20E-02
megakaryocyte differentiation(GO:0030219)	42.86	9.28	1.20E-02
germinal center formation(GO:0002467)	42.86	9.28	1.20E-02
positive regulation of metabolic process(GO:0009893)	4.01	9.00	1.20E-02
anatomical structure morphogenesis(GO:0009653)	4.44	8.76	1.20E-02
peptidyl-tyrosine phosphorylation(GO:0018108)	14.29	8.66	1.20E-02
Peyer's patch development(GO:0048541)	37.50	8.64	1.20E-02
regulation of establishment or maintenance of cell polarity(GO:0032878)	37.50	8.64	1.20E-02
luteinization(GO:0001553)	37.50	8.64	1.20E-02
positive regulation of catalytic activity(GO:0043085)	5.03	8.47	1.20E-02
positive regulation of B cell activation(GO:0050871)	15.09	8.44	1.20E-02
positive regulation of cell proliferation(GO:0008284)	5.34	8.14	1.20E-02
serotonin transport(GO:0006837)	33.33	8.11	1.20E-02
citrulline metabolic process(GO:0000052)	33.33	8.11	1.20E-02
negative regulation of coagulation(GO:0050819)	20.83	8.06	1.20E-02
cellular response to drug(GO:0035690)	25.00	7.99	1.20E-02
transforming growth factor beta receptor signaling pathway(GO:0007179)	13.11	7.75	1.20E-02
regulation of biological quality(GO:0065008)	3.59	7.66	1.20E-02
sexual reproduction(GO:0019953)	30.00	7.65	1.20E-02
succinate metabolic process(GO:0006105)	30.00	7.65	1.20E-02
natural killer cell differentiation(GO:0001779)	30.00	7.65	1.20E-02

circulatory system process(GO:0003013)	10.75	7.64	1.20E-02
positive regulation of vasoconstriction(GO:0045907)	18.52	7.53	1.20E-02
negative regulation of ossification(GO:0030279)	18.52	7.53	1.20E-02
positive regulation of cell communication(GO:0010647)	4.87	7.51	1.20E-02
positive regulation of signaling(GO:0023056)	4.81	7.43	1.20E-02
lipid storage(GO:0019915)	21.05	7.25	1.20E-02
response to retinoic acid(GO:0032526)	10.59	7.18	1.20E-02
regulation of cellular component organization(GO:0051128)	3.95	7.11	1.20E-02
negative regulation of myeloid leukocyte differentiation(GO:0002762)	16.67	7.08	1.20E-02
regulation of nucleotide biosynthetic process(GO:0030808)	8.87	7.05	1.20E-02
vesicle fusion(GO:0006906)	20.00	7.04	1.20E-02
phagocytosis(GO:0006909)	13.95	6.97	1.20E-02
small GTPase mediated signal transduction(GO:0007264)	6.21	6.96	1.20E-02
gland development(GO:0048732)	8.66	6.94	1.20E-02
arginine metabolic process(GO:0006525)	25.00	6.92	1.20E-02
regulation of muscle system process(GO:0090257)	9.38	6.62	1.20E-02
regulation of cell adhesion(GO:0030155)	6.44	6.53	1.20E-02
regulation of survival gene product expression(GO:0045884)	17.39	6.49	1.20E-02
endoderm formation(GO:0001706)	21.43	6.34	1.20E-02
regulation of tissue remodeling(GO:0034103)	13.51	6.24	1.20E-02
blood vessel development(GO:0001568)	9.09	6.11	1.20E-02
erythrocyte development(GO:0048821)	20.00	6.10	1.20E-02
female pregnancy(GO:0007565)	11.11	6.04	1.20E-02
response to estrogen stimulus(GO:0043627)	6.67	5.99	1.20E-02
regulation of fat cell differentiation(GO:0045598)	10.91	5.97	1.20E-02
vasculature development(GO:0001944)	12.50	5.94	1.20E-02
leukocyte migration(GO:0050900)	8.70	5.93	1.20E-02
negative regulation of peptidyl-tyrosine phosphorylation(GO:0050732)	18.75	5.87	1.20E-02
negative regulation of striated muscle tissue development(GO:0045843)	18.75	5.87	1.20E-02
wound healing(GO:0042060)	9.33	5.82	1.20E-02
regulation of vasodilation(GO:0042312)	14.29	5.77	1.20E-02
inner ear development(GO:0048839)	10.34	5.77	1.20E-02
heart development(GO:0007507)	6.32	5.75	1.20E-02
cell adhesion(GO:0007155)	4.08	5.68	1.20E-02
cardiac cell differentiation(GO:0035051)	11.36	5.59	1.20E-02
second-messenger-mediated signaling(GO:0019932)	6.06	5.56	1.20E-02
response to hypoxia(GO:0001666)	5.75	5.55	1.20E-02
MAPKKK cascade(GO:0000165)	7.92	5.55	1.20E-02
erythrocyte differentiation(GO:0030218)	11.11	5.51	1.20E-02
B cell differentiation(GO:0030183)	11.11	5.51	1.20E-02
regulation of caspase activity(GO:0043281)	6.71	5.49	1.20E-02
receptor-mediated endocytosis(GO:0006898)	9.52	5.46	1.20E-02
regulation of ion homeostasis(GO:2000021)	7.69	5.43	1.20E-02
transmembrane receptor protein tyrosine kinase signaling pathway(GO:0007169)	5.46	5.32	1.20E-02

SMAD protein signal transduction(GO:0060395)	15.79	5.31	1.20E-02
embryo development(GO:0009790)	4.82	5.30	1.20E-02
positive regulation of transport(GO:0051050)	4.38	5.29	1.20E-02
transcription from RNA polymerase II promoter(GO:0006366)	6.82	5.27	1.20E-02
regulation of calcium ion transport(GO:0051924)	6.67	5.18	1.20E-02
embryonic heart tube development(GO:0035050)	15.00	5.15	1.20E-02
negative regulation of transcription from RNA polymerase II promoter(GO:0000122)	4.13	5.13	1.20E-02
positive regulation of mitotic cell cycle(GO:0045931)	11.76	5.11	1.20E-02
regulation of binding(GO:0051098)	6.02	5.05	1.20E-02
regulation of bone mineralization(GO:0030500)	9.62	5.01	1.20E-02
regulation of actin filament-based process(GO:0032970)	6.34	4.98	1.20E-02
cellular response to lipopolysaccharide(GO:0071222)	8.22	4.93	1.20E-02
negative regulation of signaling(GO:0023057)	4.02	4.85	1.20E-02
negative regulation of cell communication(GO:0010648)	4.01	4.84	1.20E-02
actin filament-based process(GO:0030029)	4.83	4.78	1.20E-02
positive regulation of immune effector process(GO:0002699)	6.86	4.67	1.20E-02
divalent metal ion transport(GO:0070838)	5.07	4.58	1.20E-02
response to mechanical stimulus(GO:0009612)	6.11	4.56	1.20E-02
regulation of cytokine production(GO:0001817)	4.35	4.33	1.20E-02
regulation of cellular component biogenesis(GO:0044087)	4.53	4.32	1.20E-02
positive regulation of apoptosis(GO:0043065)	3.62	4.09	1.20E-02
developmental maturation(GO:0021700)	5.74	4.05	1.20E-02
response to metal ion(GO:0010038)	4.40	4.02	1.20E-02
aging(GO:0007568)	4.89	4.01	1.20E-02
tissue development(GO:0009888)	3.52	3.97	1.20E-02
secretion(GO:0046903)	3.69	3.51	1.20E-02
lymph vessel development(GO:0001945)	27.27	7.26	2.18E-02
phagocytosis & engulfment(GO:0006911)	21.43	6.34	2.18E-02
negative regulation of epithelial cell differentiation(GO:0030857)	18.75	5.87	2.18E-02
cAMP biosynthetic process(GO:0006171)	18.75	5.87	2.18E-02
regulation of cyclase activity(GO:0031279)	6.82	4.30	2.18E-02
regulation of multicellular organism growth(GO:0040014)	7.46	4.20	2.18E-02
regulation of lyase activity(GO:0051339)	6.59	4.19	2.18E-02
positive regulation of neurogenesis(GO:0050769)	5.51	3.92	2.18E-02
regulation of protein transport(GO:0051223)	4.02	3.32	2.18E-02
cell proliferation(GO:0008283)	3.29	3.06	2.18E-02
regulation of endothelial cell differentiation(GO:0045601)	21.43	6.34	3.09E-02
cell fate determination(GO:0001709)	12.50	5.31	3.09E-02
hemopoietic progenitor cell differentiation(GO:0002244)	12.00	4.48	3.09E-02
regulation of leukocyte mediated immunity(GO:0002703)	5.94	3.86	3.09E-02
cell fate commitment(GO:0045165)	5.26	3.49	3.09E-02
neuron differentiation(GO:0030182)	4.26	3.31	3.09E-02
signaling(GO:0023052)	3.70	3.21	3.09E-02
osteoclast differentiation(GO:0030316)	15.00	5.15	3.89E-02

regulation of alpha-beta T cell activation(GO:0046634)	8.77	4.71	3.89E-02
cellular response to hydrogen peroxide(GO:0070301)	10.26	4.68	3.89E-02
myeloid leukocyte activation(GO:0002274)	8.20	4.49	3.89E-02
membrane docking(GO:0022406)	11.11	4.27	3.89E-02
peptidyl-tyrosine dephosphorylation(GO:0035335)	8.33	4.06	3.89E-02
regulation of intracellular transport(GO:0032386)	4.90	3.54	3.89E-02
cellular response to hormone stimulus(GO:0032870)	3.68	3.18	3.89E-02
glutamine family amino acid catabolic process(GO:0009065)	17.65	5.67	4.66E-02
response to ethanol(GO:0045471)	5.45	3.60	4.66E-02

Table A.9: Significantly overrepresented, pruned, cellular component GO terms from the genes upregulated in the endocardium. Terms are sorted by ascending adjusted p-value and then by descending Z-score.

Ontology Name (Ontology-ID)	Percent Changed	Z-Score	Adjusted p-value
plasma membrane(GO:0005886)	3.25	8.90	1.20E-02
cell periphery(GO:0071944)	21.74	8.26	1.20E-02
cell surface(GO:0009986)	6.15	7.65	1.20E-02
plasma membrane part(GO:0044459)	3.24	7.12	1.20E-02
acrosomal membrane(GO:0002080)	25.00	6.92	1.20E-02
transport vesicle(GO:0030133)	10.71	5.90	1.20E-02
extracellular region part(GO:0044421)	3.26	5.26	1.20E-02
cell projection(GO:0042995)	3.21	5.01	1.20E-02
cell fraction(GO:0000267)	2.66	3.70	1.20E-02
intrinsic to Golgi membrane(GO:0031228)	11.11	4.93	2.18E-02
extrinsic to membrane(GO:0019898)	5.88	4.14	2.18E-02
Golgi membrane(GO:0000139)	3.12	2.86	3.89E-02
cell cortex part(GO:0044448)	5.95	3.53	4.66E-02

Table A.10: Significantly overrepresented, pruned, molecular function GO terms from the genes upregulated in the endocardium. Terms are sorted by ascending adjusted p-value and then by descending Z-score.

Ontology Name (Ontology-ID)	Percent Changed	Z-Score	Adjusted p-value
RNA polymerase II regulatory region sequence-specific DNA binding(GO:0000977)	15.87	9.73	1.20E-02
protein binding(GO:0005515)	2.32	9.73	1.20E-02
pattern binding(GO:0001871)	7.23	6.37	1.20E-02
calcium ion binding(GO:0005509)	4.35	6.17	1.20E-02
core promoter sequence-specific DNA binding(GO:0001046)	15.38	6.03	1.20E-02
sequence-specific DNA binding transcription factor activity(GO:0003700)	3.54	5.26	1.20E-02
guanyl-nucleotide exchange factor activity(GO:0005085)	6.33	5.25	1.20E-02
enzyme activator activity(GO:0008047)	4.49	4.80	1.20E-02
chromatin binding(GO:0003682)	4.38	4.01	1.20E-02
receptor signaling protein activity(GO:0005057)	6.41	3.74	1.20E-02
vascular endothelial growth factor receptor activity(GO:0005021)	37.50	8.64	2.18E-02
hydrolase activity & acting on carbon-nitrogen (but not peptide) bonds & in linear amidines(GO:0016813)	27.27	7.26	2.18E-02
sialyltransferase activity(GO:0008373)	15.79	5.31	3.09E-02
cAMP binding(GO:0030552)	16.67	5.48	3.89E-02
protein tyrosine phosphatase activity(GO:0004725)	6.19	3.99	3.89E-02

Table A.11: Significantly overrepresented, pruned, biological process GO terms from the genes upregulated in the endothelium. Terms are sorted by ascending adjusted p-value and then by descending Z-score.

Ontology Name(Ontology-ID)	Percent Changed	Z-Score	Adjusted p-value
cell-cell adhesion(GO:0016337)	14.44	15.26	2.07E-02
anatomical structure development(GO:0048856)	5.84	13.59	2.07E-02
embryonic pattern specification(GO:0009880)	25.49	12.40	2.07E-02
forebrain anterior/posterior pattern formation(GO:0021797)	66.67	11.66	2.07E-02
anatomical structure morphogenesis(GO:0009653)	6.52	11.65	2.07E-02
axon guidance(GO:0007411)	15.75	11.51	2.07E-02
axis specification(GO:0009798)	20.00	11.15	2.07E-02
multicellular organismal development(GO:0007275)	6.83	11.02	2.07E-02
tube formation(GO:0035148)	16.16	10.46	2.07E-02
cellular developmental process(GO:0048869)	4.87	9.78	2.07E-02
negative regulation of cell differentiation(GO:0045596)	8.13	8.89	2.07E-02
Wnt receptor signaling pathway involved in somitogenesis(GO:0090244)	50.00	8.66	2.07E-02
response to vitamin(GO:0033273)	10.39	7.78	2.07E-02
regulation of nervous system development(GO:0051960)	6.82	7.68	2.07E-02
hair cycle process(GO:0022405)	14.29	7.63	2.07E-02
regulation of transcription from RNA polymerase II promoter(GO:0006357)	4.98	7.55	2.07E-02
cell migration(GO:0016477)	6.70	7.44	2.07E-02
positive regulation of fat cell differentiation(GO:0045600)	23.81	7.38	2.07E-02
negative chemotaxis(GO:0050919)	28.57	7.34	2.07E-02
regulation of cell fate commitment(GO:0010453)	22.73	7.18	2.07E-02
primitive streak formation(GO:0090009)	33.33	6.93	2.07E-02
development of primary sexual characteristics(GO:0045137)	33.33	6.93	2.07E-02
cell communication(GO:0007154)	6.14	6.88	2.07E-02
cellular component maintenance(GO:0043954)	17.65	6.76	2.07E-02
extracellular structure organization(GO:0043062)	8.33	6.59	2.07E-02
regulation of cell adhesion(GO:0030155)	7.73	6.58	2.07E-02
regulation of embryonic development(GO:0045995)	12.33	6.57	2.07E-02
mammary gland involution(GO:0060056)	30.00	6.53	2.07E-02
negative regulation of JUN kinase activity(GO:0043508)	30.00	6.53	2.07E-02
regulation of cell proliferation(GO:0042127)	4.55	6.31	2.07E-02
neg. reg. of TM receptor protein Ser/Thr kinase signaling pathway(GO:0090101)	12.50	6.25	2.07E-02
dorsal/ventral pattern formation(GO:0009953)	12.31	6.18	2.07E-02
positive regulation of Notch signaling pathway(GO:0045747)	21.05	6.14	2.07E-02
positive regulation of transcription & DNA-dependent(GO:0045893)	4.57	6.11	2.07E-02
segmentation(GO:0035282)	13.21	6.06	2.07E-02
regulation of smoothened signaling pathway(GO:0008589)	14.63	6.00	2.07E-02
regulation of Wnt receptor signaling pathway(GO:0030111)	8.00	5.52	2.07E-02
regulation of cell migration(GO:0030334)	5.88	5.46	2.07E-02
negative regulation of locomotion(GO:0040013)	8.85	5.45	2.07E-02

regulation of calcium ion-dependent exocytosis(GO:0017158)	16.00	5.19	2.07E-02
canonical Wnt receptor signaling pathway(GO:0060070)	9.52	5.15	2.07E-02
negative regulation of transcription & DNA-dependent(GO:0045892)	4.38	5.12	2.07E-02
developmental growth involved in morphogenesis(GO:0060560)	11.32	5.05	2.07E-02
establishment or maintenance of cell polarity(GO:0007163)	9.86	4.94	2.07E-02
cell proliferation(GO:0008283)	5.21	4.70	2.07E-02
locomotory behavior(GO:0007626)	6.67	4.32	2.07E-02
response to endogenous stimulus(GO:0009719)	3.91	4.19	2.07E-02
cellular response to chemical stimulus(GO:0070887)	3.70	4.14	2.07E-02
response to wounding(GO:0009611)	4.48	3.85	2.07E-02
sensory perception of sound(GO:0007605)	7.14	3.83	2.07E-02
embryo development(GO:0009790)	4.52	3.55	2.07E-02
cellular process involved in reproduction(GO:0048610)	4.82	3.42	2.07E-02
regulation of biological quality(GO:0065008)	2.96	3.28	2.07E-02
detection of mechanical stimulus involved in sensory perception of sound(GO:0050910)	25.00	5.88	3.84E-02
regulation of cell division(GO:0051302)	8.22	3.98	3.84E-02

Table A.12: Significantly overrepresented, pruned, cellular component GO terms from the genes upregulated in the endothelium. Terms are sorted by ascending adjusted p-value and then by descending Z-score.

Ontology Name(Ontology-ID)	Percent Changed	Z-Score	Adjusted p-value
extracellular matrix(GO:0031012)	9.94	10.87	2.07E-02
cell-cell junction(GO:0005911)	9.06	8.63	2.07E-02
extracellular region(GO:0005576)	4.67	8.30	2.07E-02
fibrillar collagen(GO:0005583)	33.33	8.00	2.07E-02
extracellular space(GO:0005615)	5.60	7.80	2.07E-02
cell surface(GO:0009986)	7.26	7.54	2.07E-02
lateral plasma membrane(GO:0016328)	20.69	7.44	2.07E-02
axon(GO:0030424)	8.37	6.82	2.07E-02
apical plasma membrane(GO:0016324)	7.73	6.40	2.07E-02
plasma membrane(GO:0005886)	3.36	6.18	2.07E-02
intrinsic to plasma membrane(GO:0031226)	5.34	5.67	2.07E-02
dendrite(GO:0030425)	6.51	5.53	2.07E-02
anchoring junction(GO:0070161)	7.11	5.41	2.07E-02
basement membrane(GO:0005604)	9.64	5.20	2.07E-02
cell projection part(GO:0044463)	4.72	4.95	2.07E-02
site of polarized growth(GO:0030427)	8.51	4.73	2.07E-02
synapse(GO:0045202)	5.04	4.42	2.07E-02
filopodium(GO:0030175)	10.64	4.41	2.07E-02
basolateral plasma membrane(GO:0016323)	6.71	4.34	2.07E-02
synapse part(GO:0044456)	4.28	3.43	2.07E-02
perinuclear region of cytoplasm(GO:0048471)	4.06	3.30	2.07E-02

Table A.13: Significantly overrepresented, pruned, molecular function GO terms from the genes upregulated in the endothelium. Terms are sorted by ascending adjusted p-value and then by descending Z-score.

Ontology Name(Ontology-ID)	Percent Changed	Z-Score	Adjusted p-value
chemorepellent activity(GO:0045499)	60.00	9.55	2.07E-02
extracellular matrix binding(GO:0050840)	21.05	8.69	2.07E-02
transmembrane-ephrin receptor activity(GO:0005005)	50.00	8.66	2.07E-02
pattern binding(GO:0001871)	10.84	8.51	2.07E-02
calcium ion binding(GO:0005509)	6.52	8.46	2.07E-02
inorganic anion exchanger activity(GO:0005452)	33.33	8.00	2.07E-02
Wnt receptor activity(GO:0042813)	25.00	7.60	2.07E-02
axon guidance receptor activity(GO:0008046)	37.50	7.40	2.07E-02
heparan sulfate proteoglycan binding(GO:0043395)	28.57	7.34	2.07E-02
Wnt-protein binding(GO:0017147)	20.00	6.66	2.07E-02
PDZ domain binding(GO:0030165)	7.92	4.47	2.07E-02
sequence-specific DNA binding RNA polymerase II transcription factor activity(GO:0000981)	6.36	4.34	2.07E-02
receptor binding(GO:0005102)	3.23	3.33	2.07E-02
identical protein binding(GO:0042802)	3.27	3.05	2.07E-02
growth factor binding(GO:0019838)	7.14	4.10	3.84E-02

A.6 TF Distribution near Transcription Start Sites

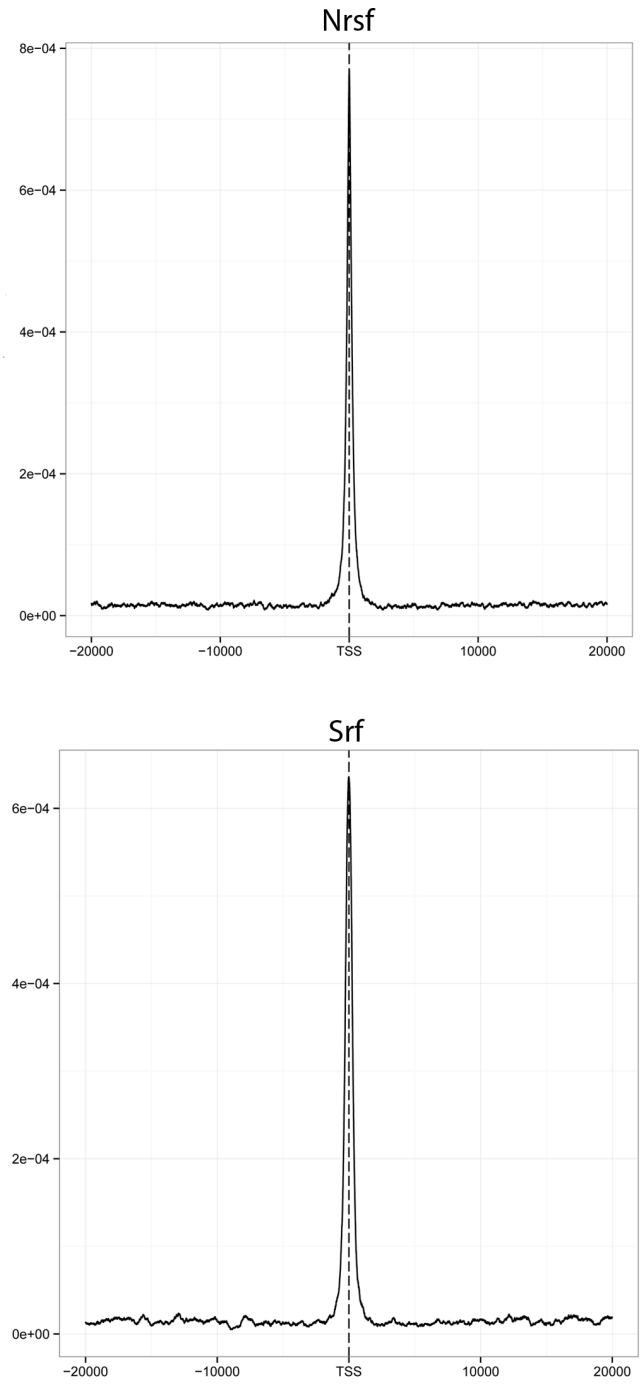


Figure A.1: Distribution of transcription factors Nrsf and Srf in the vicinity of TSS. Plots produced with publicly available data. The x axis denotes distance from the TSS and the y peak density.

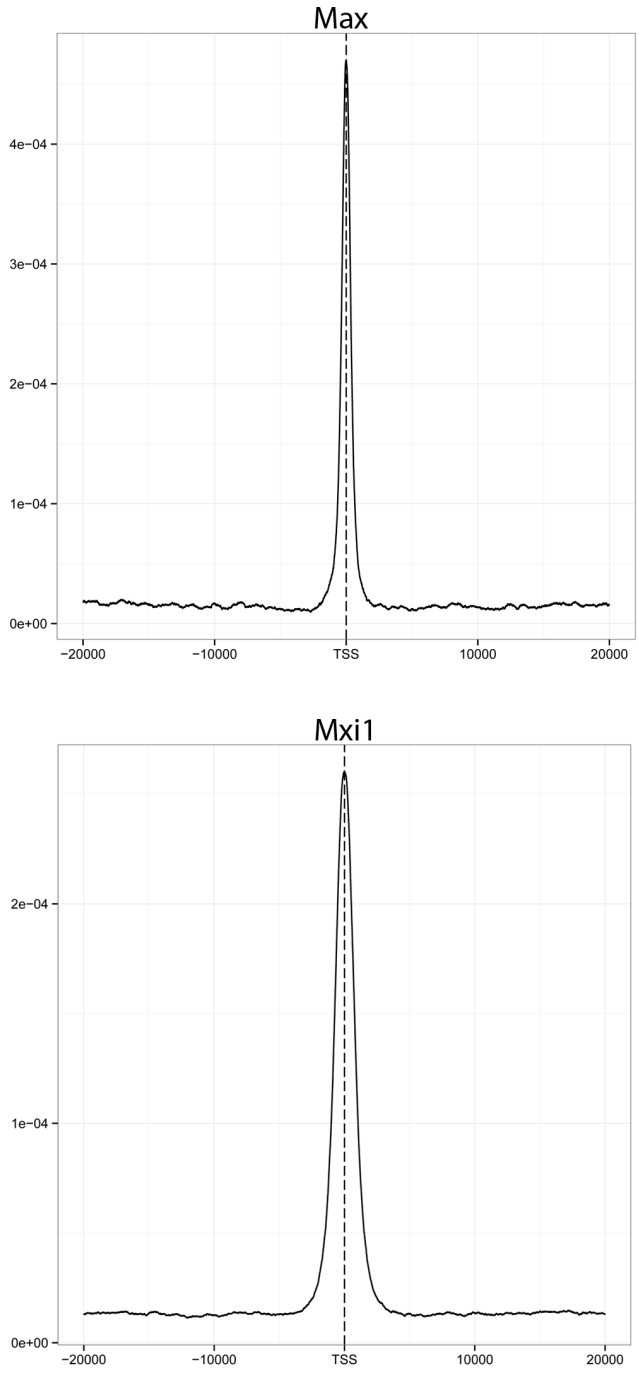


Figure A.2: Distribution of transcription factors Max and Mxi1 in the vicinity of TSS. Plots produced with publicly available data. The x axis denotes distance from the TSS and the y peak density.

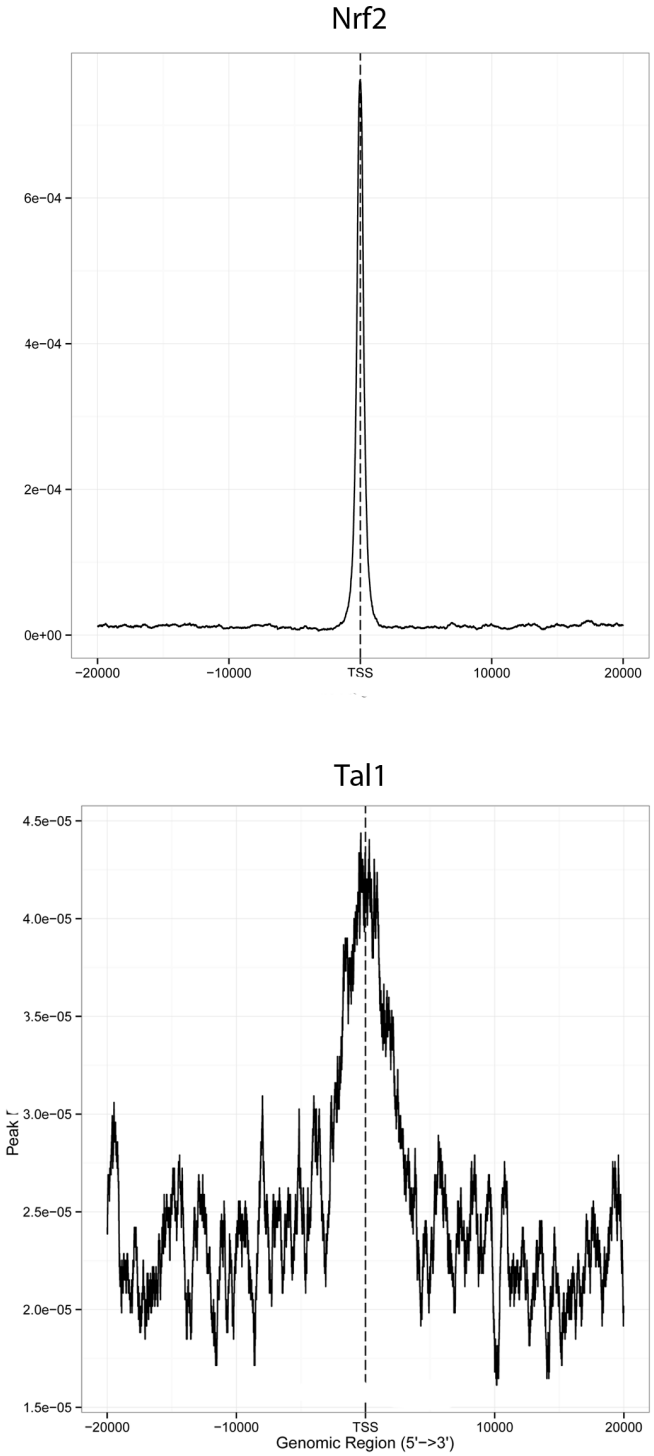


Figure A.3: Distribution of transcription factors Nrf2 and Tal1 in the vicinity of TSS. Plots produced with publicly available data. The x axis denotes distance from the TSS and the y peak density.

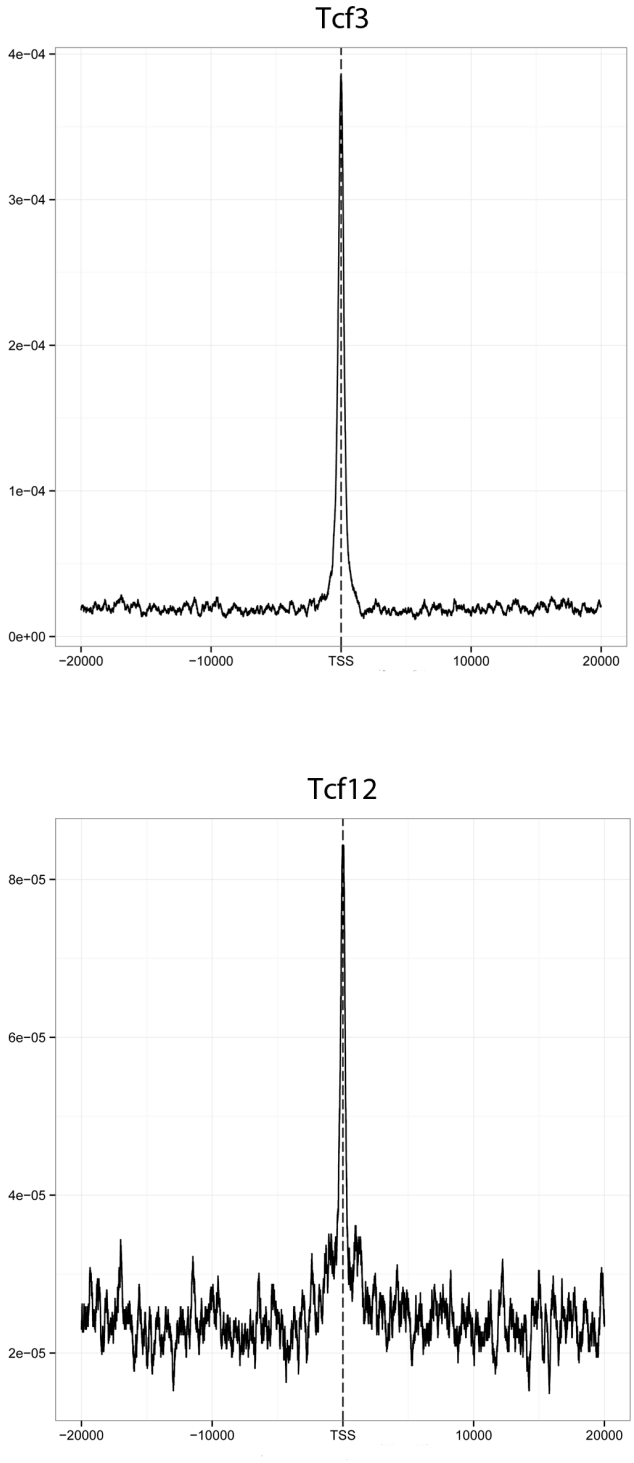


Figure A.4: Distribution of transcription factors Tcf3 and Tcf12 in the vicinity of TSS. Plots produced with publicly available data. The x axis denotes distance from the TSS and the y peak density.

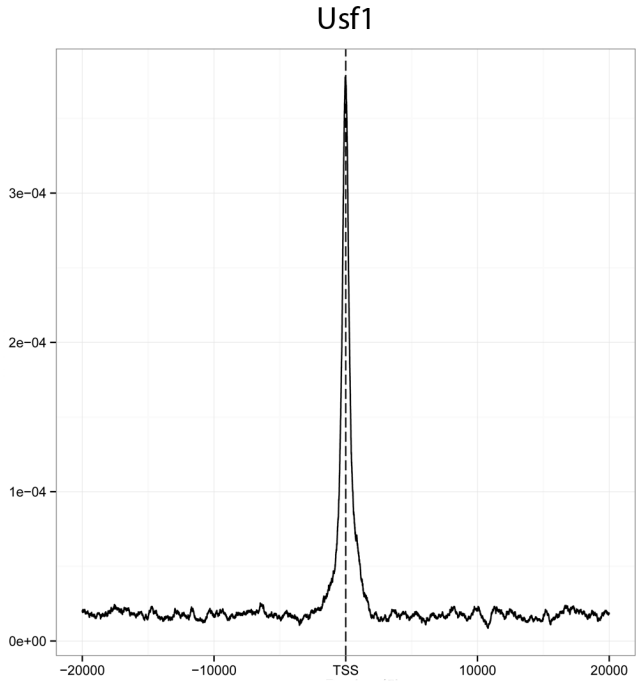


Figure A.5: Distribution of transcription factor Usf1 in the vicinity of TSS. Plots produced with publicly available data. The x axis denotes distance from the TSS and the y peak density.

A.7 Overlap of Differentially Methylated Genomic Regions and Differentially Expressed Genes

Table A.14: Overlap of Differentially Methylated Genomic Regions and Differentially Expressed Genes

CGI Locus	EC CGI methylation	Gene Locus	Gene Name	log2(FC) in EC
chr2:152896544-152896824	hyper	chr2:152891690-152907471	Ccm2l	3.05366
chr7:117952877-117953693	hyper	chr7:117994120-118006467	Lyve1	2.86907
chr7:52887987-52888323	hyper	chr7:52882906-52894462	Rasip1	2.76976
chr11:96440246-96440681	hyper	chr11:96325904-96620936	Skap1	2.7605
chr11:96338361-96338945	hyper	chr11:96325904-96620936	Skap1	2.7605
chr8:73928811-73929522	hyper	chr8:73908172-73919704	Ushbp1	2.66224
chr4:118108001-118108883	hyper	chr4:118143795-118162454	Tie1	2.61175
chr4:118190073-118190950	hyper	chr4:118143795-118162454	Tie1	2.61175
chr10:74816185-74817404	hypo	chr10:74779687-74797533	Adora2a	2.33696
chr4:114751244-114752324	hyper	chr4:114729365-114744360	Tal1	2.30528
chr10:87323389-87323840	hyper	chr10:87321800-87399792	Igf1	2.25402
chr4:114751244-114752324	hyper	chr4:114729365-114744360	Tal1	2.24749
chr11:70028833-70029223	hyper	chr11:70037628-70043300	Bcl6b	2.23097
chr4:114751244-114752324	hyper	chr4:114729365-114744360	Tal1	2.22898
chr2:103843358-103843795	hyper	chr2:103798151-103822035	Lmo2	2.11793
chr11:49403109-49403751	hyper	chr11:49423180-49466241	Flt4	2.06644
chr6:88154810-88156434	hyper	chr6:88148657-88157026	Gata2	2.03072
chr19:37507291-37507847	hyper	chr19:37509330-37515221	Hhex	1.90728
chr5:150057323-150058420	hypo	chr5:150076633-150099623	Alox5ap	1.83045
chr12:112661520-112662384	hyper	chr12:112680871-112693229	Tnfaip2	1.71677
chr2:103843358-103843795	hyper	chr2:103798151-103822035	Lmo2	1.71509
chr6:115912422-115912546	hyper	chr6:115904828-115945023	Plxnd1	1.6793
chr8:107860607-107860809	hyper	chr8:107813823-107820136	Exoc3l	1.677
chr12:74894209-74894870	hyper	chr12:74686027-74879171	Prkch	1.64118
chr3:79683120-79683489	hyper	chr3:79689851-79750200	Fam198b	1.60056
chr14:56380507-56380930	hyper	chr14:56387928-56402856	Adcy4	1.59697
chr14:56434090-56434667	hyper	chr14:56387928-56402856	Adcy4	1.59697
chr7:31250804-31251368	hyper	chr7:31198806-31202598	Tyrobp	1.55592
chr16:92475463-92476098	hyper	chr16:92498391-92541486	Clic6	1.5264
chr16:92468411-92468768	hyper	chr16:92498391-92541486	Clic6	1.5264
chr7:20055181-20055460	hyper	chr7:20080256-20082313	-	1.52238
chr18:61228464-61228781	hyper	chr18:61265225-61290793	Csf1r	1.3623
chr18:61225022-61225835	hyper	chr18:61265225-61290793	Csf1r	1.3623
chr1:187308178-187308821	hyper	chr1:187270994-187292640	Slc30a10	1.30679
chr14:55330531-55330769	hyper	chr14:55341051-55400723	Slc7a8	1.14997
chr5:36083518-36083985	hyper	chr5:36039828-36071925	Sh3tc1	1.10348
chr19:37765321-37765560	hyper	chr19:37624907-37758502	Exoc6	1.09551
chr19:37762177-37762475	hyper	chr19:37624907-37758502	Exoc6	1.09551
chr7:105313648-105314322	hyper	chr7:105270074-105333309	Capn5	1.07571
chr7:105313648-105314322	hyper	chr7:105199563-105268003	Myo7a	1.06466

chr5:122278315-122278607	hyper	chr5:122265469-122286810	Sh2b3	1.06012
chr6:142845417-142846067	hyper	chr6:142762750-142912972	St8sia1	1.03431
chr8:124859685-124859999	hyper	chr8:124806040-124861147	Zfpm1	1.03023
chr8:124793057-124793175	hyper	chr8:124806040-124861147	Zfpm1	1.03023
chr8:124860571-124861615	hyper	chr8:124806040-124861147	Zfpm1	1.03023
chr8:124820289-124820793	hypo	chr8:124806040-124861147	Zfpm1	1.03023
chr8:124849801-124850807	hypo	chr8:124806040-124861147	Zfpm1	1.03023
chr8:124845529-124845986	hypo	chr8:124806040-124861147	Zfpm1	1.03023
chr2:158625885-158627235	hyper	chr2:158492468-158592070	Ppp1r16b	1.00864
chr15:80416857-80417546	hyper	chr15:80453912-80483450	Grap2	1.00605
chr1:167565599-167566481	hyper	chr1:167579075-167638225	Rcsd1	0.943436
chr8:108090552-108090881	hyper	chr8:108129128-108146118	Fam65a	0.881099
chr11:106468008-106468818	hyper	chr11:106515531-106585695	Pecam1	0.881004
chr19:37765321-37765560	hyper	chr19:37772297-37776026	Cyp26a1	0.848688
chr19:37762177-37762475	hyper	chr19:37772297-37776026	Cyp26a1	0.848688
chr19:37814091-37814949	hypo	chr19:37772297-37776026	Cyp26a1	0.848688
chr7:104473728-104473923	hyper	chr7:104230260-104457461	Gab2	0.790224
chr5:117627706-117628918	hyper	chr5:117570137-117725107	Taok3	0.747814
chr5:117690695-117691301	hyper	chr5:117570137-117725107	Taok3	0.747814
chr12:113027316-113027845	hyper	chr12:113066668-113146266	Ppp1r13b	0.729111
chr19:5751116-5751276	hyper	chr19:5707373-5726317	Ehbp1l1	0.716896
chr7:86503189-86504238	hypo	chr7:86537223-86611159	Polg	0.662579
chr11:100819860-100820506	hyper	chr11:100748123-100800825	Stat3	0.65769
chr5:129099014-129099686	hyper	chr5:129106980-129109968	Fzd10	0.654989
chr19:5751116-5751276	hyper	chr19:5689130-5702864	Map3k11	0.640204
chr19:57064818-57065598	hyper	chr19:57107753-57290522	Ablim1	0.614848
chr19:57216314-57217562	hyper	chr19:57107753-57290522	Ablim1	0.614848
chr13:38033390-38033995	hyper	chr13:37917906-38043929	Rreb1	0.614741
chr13:38022600-38023145	hyper	chr13:37917906-38043929	Rreb1	0.614741
chr7:26430147-26430915	hyper	chr7:26472020-26490015	Tgfb1	0.601568
chr7:26489381-26490333	hyper	chr7:26472020-26490015	Tgfb1	0.601568
chr9:63801417-63801870	hyper	chr9:63800882-63869866	Smad6	0.58375
chr9:63775068-63775725	hyper	chr9:63800882-63869866	Smad6	0.58375
chr9:63766729-63767498	hypo	chr9:63800882-63869866	Smad6	0.58375
chr11:100819860-100820506	hyper	chr11:100818050-100831931	Ptrf	0.57786
chr4:131478296-131478516	hyper	chr4:131477064-131631228	Epb4.1	0.550298
chr17:35154731-35154891	hyper	chr17:35195979-35199044	Ddah2	0.546457
chr17:6940931-6941564	hyper	chr17:6942479-6987129	Ezr	-0.502139
chr17:6988252-6988957	hyper	chr17:6942479-6987129	Ezr	-0.502139
chr8:46136332-46136765	hyper	chr8:46020611-46137611	Fat1	-0.540317
chr4:136312041-136312761	hyper	chr4:136203513-136391850	Ephb2	-0.578212
chr7:56887804-56888417	hyper	chr7:56214442-56865458	Nav2	-0.614748
chr4:129807866-129808438	hyper	chr4:129842843-129852366	Tinagl1	-0.622873
chr4:129825701-129826354	hyper	chr4:129842843-129852366	Tinagl1	-0.622873

chr18:20759818-20760284	hyper	chr18:20716616-20763027	Dsg2	-0.634473
chr12:113958407-113959505	hyper	chr12:113960384-113984802	Cep170b	-0.642405
chr2:158222992-158223588	hyper	chr2:158201373-158211881	Snhg11	-0.643023
chr8:73269722-73269943	hyper	chr8:73286671-73291611	Ifi30	-0.651521
chr8:73309871-73310263	hyper	chr8:73286671-73291611	Ifi30	-0.651521
chr8:73334834-73335143	hyper	chr8:73286671-73291611	Ifi30	-0.651521
chr9:14050690-14051104	hyper	chr9:14080744-14137524	Sesn3	-0.690743
chr1:34093309-34093859	hyper	chr1:34068669-34365497	Dst	-0.712324
chr4:126002087-126002467	hyper	chr4:125964037-125991574	Col8a2	-0.755375
chr4:125923594-125924274	hyper	chr4:125964037-125991574	Col8a2	-0.755375
chr7:148211594-148212613	hyper	chr7:148247172-148258018	B4galnt4	-0.834077
chr7:148274750-148274879	hyper	chr7:148247172-148258018	B4galnt4	-0.834077
chr12:118397824-118398305	hyper	chr12:117724192-118575485	Ptprn2	-0.844585
chr12:118266635-118267509	hyper	chr12:117724192-118575485	Ptprn2	-0.844585
chr8:124793057-124793175	hyper	chr8:124783835-124796514	-	-0.849553
chr8:124820289-124820793	hypo	chr8:124783835-124796514	-	-0.849553
chr8:124845529-124845986	hypo	chr8:124783835-124796514	-	-0.849553
chr3:121787724-121788914	hyper	chr3:121747377-121882979	Abca4	-0.865933
chr5:115894491-115895052	hyper	chr5:115879693-115905695	Msi1	-0.879537
chr14:61623709-61624298	hyper	chr14:61582670-61665692	Tnfrsf19	-0.882696
chr7:148192086-148193168	hyper	chr7:148153327-148155726	Ifitm1	-0.882976
chr3:84074010-84074594	hyper	chr3:83964360-84108697	Trim2	-1.1074
chr7:28066903-28067263	hyper	chr7:28090159-28122631	Ltbp4	-1.11708
chr5:44490853-44491527	hyper	chr5:44384860-44492975	Prom1	-1.12478
chr11:35610726-35611148	hyper	chr11:35651913-35793591	Wwc1	-1.25767
chr7:4434397-4434994	hyper	chr7:4469909-4484044	Dnaaf3	-1.2718
chr5:141088065-141088599	hyper	chr5:141083294-141091499	Lfng	-1.29608
chr4:116502237-116503120	hyper	chr4:116550006-116661710	Zswim5	-1.32418
chr11:102217890-102218476	hyper	chr11:102210133-102226595	Slc4a1	-1.34342
chr11:102639062-102639563	hyper	chr11:102622752-102641576	Adam11	-1.34449
chr14:122852090-122852827	hyper	chr14:122874605-122879550	Zic2	-1.39138
chr7:51857056-51858580	hyper	chr7:51861172-51926213	Myh14	-1.42359
chr7:29997949-29998493	hyper	chr7:30041348-30066996	Spint2	-1.43921
chr10:19861332-19862037	hyper	chr10:19868725-20001396	Map7	-1.49083
chr13:24613355-24613926	hyper	chr13:24419288-24569154	Cmah	-1.54945
chr3:108187823-108188389	hyper	chr3:108193765-108218412	Celsr2	-1.57786
chr17:15519420-15519960	hyper	chr17:15504317-15512787	Dll1	-1.75381
chr10:59570816-59571512	hyper	chr10:59569004-59597899	Spock2	-1.81204
chr4:124344632-124345709	hyper	chr4:124334888-124337899	Pou3f1	-1.83496
chr4:125212061-125213486	hyper	chr4:125168074-125391417	Grik3	-1.86844
chr4:125222430-125222877	hyper	chr4:125168074-125391417	Grik3	-1.86844
chr1:138696965-138697319	hyper	chr1:138740160-138857025	Nr5a2	-1.87529
chr2:147849058-147849438	hyper	chr2:147868613-147872705	Foxa2	-1.87655
chr4:45801580-45802815	hypo	chr4:45822378-45839699	Igfbpl1	-2.0404

chr1:182824490-182825286	hyper	chr1:182865169-182868532	Lefty1	-2.09009
chr8:109241931-109242163	hyper	chr8:109127267-109194146	Cdh1	-2.09257
chr1:72284301-72284987	hyper	chr1:72205806-72258881	Mreg	-2.09585
chr9:110817103-110817617	hyper	chr9:110842111-110848662	Tdgf1	-2.14115
chr3:34555551-34556010	hyper	chr3:34459302-34576915	Sox2	-2.17126
chr8:4206269-4206850	hyper	chr8:4209542-4217312	BC068157	-2.22997
chr5:135442964-135443614	hyper	chr5:135420992-135422804	Cldn4	-2.28976
chr11:116848981-116849900	hyper	chr11:116780176-116848258	Mgat5b	-2.33929
chr4:41044457-41044778	hyper	chr4:41039756-41045216	Aqp3	-2.3542
chr11:97445466-97446044	hyper	chr11:97370653-97436440	Srcin1	-2.46769
chr11:68246148-68246397	hyper	chr11:68022865-68200328	Ntn1	-2.48021
chr5:135442964-135443614	hyper	chr5:135462083-135477220	Cldn3	-2.60028
chr4:135846684-135847192	hyper	chr4:135803871-135830814	Tcea3	-2.75576
chr1:137114015-137114765	hyper	chr1:137150150-137155049	Elf3	-2.87026
chr7:56887804-56888417	hyper	chr7:56886868-56892205	Dbx1	-2.92245
chr15:85337985-85338351	hyper	chr15:85365866-85424138	Wnt7b	-3.09208
chr15:98626660-98627388	hyper	chr15:98620287-98624261	Wnt1	-3.09913
chr15:84961845-84962260	hyper	chr15:84895118-84962387	Smc1b	-3.21409
chr7:148274750-148274879	hyper	chr7:148287117-148303705	Ano9	-3.22434
chr12:87720288-87721157	hyper	chr12:87702066-87862578	Esrrb	-3.27954
chr15:76004856-76005224	hyper	chr15:75931917-75956986	BC024139 & Eppk1	-3.62001
chr6:122513110-122513678	hyper	chr6:122555420-122560089	Gdf3	-3.85083

Appendix B

Publications

Research

Genome-wide and parental allele-specific analysis of CTCF and cohesin DNA binding in mouse brain reveals a tissue-specific binding pattern and an association with imprinted differentially methylated regions

Adam R. Prickett,[1] Nikolaos Barkas,[1] Ruth B. McCole,[1] Siobhan Hughes, Samuele M. Amante, Reiner Schulz, and Rebecca J. Oakey[2]

Department of Medical & Molecular Genetics, King's College London, Guy's Hospital, London, SE1 9RT, United Kingdom

DNA binding factors are essential for regulating gene expression. CTCF and cohesin are DNA binding factors with central roles in chromatin organization and gene expression. We determined the sites of CTCF and cohesin binding to DNA in mouse brain, genome wide and in an allele-specific manner with high read-depth ChIP-seq. By comparing our results with existing data for mouse liver and embryonic stem (ES) cells, we investigated the tissue specificity of CTCF binding sites. ES cells have fewer unique CTCF binding sites occupied than liver and brain, consistent with a ground-state pattern of CTCF binding that is elaborated during differentiation. CTCF binding sites without the canonical consensus motif were highly tissue specific. In brain, a third of CTCF and cohesin binding sites coincide, consistent with the potential for many interactions between cohesin and CTCF but also many instances of independent action. In the context of genomic imprinting, CTCF and/or cohesin bind to a majority but not all differentially methylated regions, with preferential binding to the unmethylated parental allele. Whether the parental allele-specific methylation was established in the parental germlines or post-fertilization in the embryo is not a determinant in CTCF or cohesin binding. These findings link CTCF and cohesin with the control regions of a subset of imprinted genes, supporting the notion that imprinting control is mechanistically diverse.

[Supplemental material is available for this article.]

DNA sequences that control transcription are frequently located in the noncoding portion of the mammalian genome (The ENCODE Project Consortium 2012). These elements can act over long distances (Noonan and McCallion 2010). The identification of these control elements is important for elucidating human genetic disease since genome-wide association studies regularly point to noncoding regions as candidates in disease etiology (Manolio 2010). One of the proteins that contributes to the regulation of gene expression across the genome is CTCF (CCCTC-binding factor), a protein with 11 zinc fingers (Filippova et al. 1996) and multiple regulatory functions (Ohlsson 2001; Gaszner and Felsenfeld 2006). CTCF can act as an insulator by blocking interactions between enhancers and promoters (Bell et al. 1999), it can directly regulate chromosomal interactions (Yusufzai and Felsenfeld 2004; Hadjur et al. 2009), and it can act as an enhancer of transcription (Kuzmin et al. 2005). CTCF binds regions of DNA with high sequence specificity and is sensitive to DNA methylation, having a lower binding affinity for methylated DNA (Mukhopadhyay et al. 2004). The canonical consensus binding motif of CTCF and the sites of CTCF binding are evolutionarily conserved between mammals and birds (Martin et al. 2011; Schmidt et al. 2012). In vitro assays have shown that CTCF can use different combinations of its zinc fingers to bind to distinct DNA sequences (Filippova et al. 1996). CTCF interacts with a variety of other factors. In particular, the cohesin complex, best known for its role in mediating sister-chromatid cohesion during cell division, has been found to frequently colocalize with CTCF during interphase (Parelho et al. 2008; Rubio et al. 2008; Wendt et al. 2008; Xiao et al. 2011) with consequences for gene expression. At specific loci, cohesin is required for cell-type-specific long-range chromosomal interactions in *cis* during cellular differentiation (Hadjur et al. 2009).

Genomic imprinting refers to the parental allele-specific transcription of a subset of genes in mammals and flowering plants (Reik and Walter 2001; da Rocha et al. 2008). Roughly 140 transcripts are known to be imprinted in mammals (Schulz et al. 2008). Imprinting is controlled by epigenetic modifications that differ between the two parental genomes, including differences in DNA methylation (Li et al. 1993). Imprinted genes can occur in large, coordinately regulated clusters exemplified by the *Gnas* locus (Peters and Williamson 2007); they can form small domains such as the *Mcts2/H13* locus (Wood et al. 2008) that are comprised of only two genes (McCole and Oakey 2008), or they can be singletons like *Impact* (Hagiwara et al. 1997). In all cases, their parental allele-specific expression is ultimately due to an imprinting control region (ICR), a region of DNA that is differentially methylated between the parental alleles. The parental allele-specific methylation of a differentially methylated region (DMR) is in most cases the consequence of the sex-specific epigenetic reprogramming of the

[1]These authors contributed equally to this work.
[2]Corresponding author
E-mail rebecca.oakey@kcl.ac.uk
Article published online before print. Article, supplemental material, and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.150136.112.

23:1624–1635 © 2013, Published by Cold Spring Harbor Laboratory Press; ISSN 1088-9051/13; www.genome.org

parental germ cells (Edwards and Ferguson-Smith 2007; Bartolomei 2009). In addition, DMRs are actively protected from post-fertilization epigenetic reprogramming (Quenneville et al. 2012) and, thus, persist into adulthood. In some cases, the parental allele-specific methylation of a DMR is set up post-fertilization during early embryogenesis (somatic DMRs) (Kobayashi et al. 2012). DMRs with a direct germline origin are referred to as germline DMRs (gDMRs). Disruption of imprinted gene expression after deletion of a DMR is considered evidence that the latter functions as an ICR. There are 22 well-established gDMRs in mouse, of which 19 are maternally methylated and three are paternally methylated. Many mechanisms exist to "translate" allele-specific methylation into differential gene expression, including differential protein binding (Lewis and Reik 2006).

CTCF has long been associated with genomic imprinting due to its selective binding of the unmethylated maternal allele of the *Igf2/H19* ICR resulting in parent-of-origin-specific expression of *Igf2* and *H19* in mouse and human (Bell and Felsenfeld 2000; Hark et al. 2000; Kanduri et al. 2000; Fedoriw et al. 2004; Szabo et al. 2004). CTCF has been studied at several other imprinted loci, and it binds the unmethylated allele at the gDMRs of *Rasgrf1*, *Peg13*, *Kcnq1ot1* (Yoon et al. 2005; Fitzpatrick et al. 2007; Singh et al. 2011), and *Grb10* (Hikichi et al. 2003; Mukhopadhyay et al. 2004). CTCF-mediated regulation is postulated to be one of two major control mechanisms operating at ICRs (Lewis and Reik 2006; Kim et al. 2009). Cohesin also has been linked to imprinting through its association with CTCF at the *H19/Igf2* and *Kcnq1ot1* DMRs (Stedman et al. 2008; Lin et al. 2011), and a role for cohesin in the allele-specific organization of higher-order chromatin has been proposed (Nativio et al. 2009). Here we present the first comprehensive analysis of allele-specific CTCF and cohesin binding at all known DMRs in a single tissue, providing an unbiased assessment of the extent to which CTCF and cohesin are involved in imprinting control.

Genome-wide ChIP-seq in mouse ES cells (Chen et al. 2008; Kagey et al. 2010) and human cells (Kim et al. 2007b) has shown that CTCF and cohesin bind tens of thousands of discrete sites across the genome, and CTCF binding is enriched in and near genes, consistent with a role in the control of gene expression. Mouse embryonic stem (ES) cell data identify CTCF and cohesin binding at the gDMRs of *Peg13*, *Zim2* (*Peg3*), *Peg10*, *Grb10*, and *Mest* but not at the *H19/Igf2* ICR, even though CTCF is known to be important for imprinting regulation at this domain. Imprinting is dispensable in ES cells, where loss of imprinting frequently occurs without affecting viability in culture (Kim et al. 2007a; Rugg-Gunn et al. 2007; Frost et al. 2011). The same is true for *Dnmt1*$^{-/-}$, *Dnmt3a*$^{-/-}$, *Dnmt3b*$^{-/-}$ triple knockout mouse ES cells that consequently lack all DNA methylation imprints but yet are viable (Tsumura et al. 2006). In contrast, a differentiated tissue where imprinting plays an important role is the brain (Davies et al. 2007), and this is supported by multiple lines of evidence. Firstly, the human imprinting disorders Prader-Willi syndrome and Angelman syndrome present with behavioral and neurodevelopmental phenotypes (Cassidy et al. 2000; Lossie et al. 2001; Williams et al. 2006); secondly, of the ~140 imprinted gene transcripts in the mouse, more than 50 are expressed in brain (Wilkins 2008); thirdly, the disruption of certain mouse imprinted genes, including *Peg3* (Li et al. 1999), *Mest* (Lefebvre et al. 1998), *Nesp55* (Plagge et al. 2005), and *Grb10* (Garfield et al. 2011), results in behavioral phenotypes; finally, genome-wide allele-specific studies of transcription in mouse brain suggest that this tissue is a focus for imprinted gene expression (Gregg et al. 2010a,b; DeVeale et al. 2012).

Our analyses of CTCF and cohesin binding in mouse brain are based on ChIP-seq data of high quality and an order of magnitude higher read depth than existing data. The use of reciprocal intersubspecies hybrid mice enabled independent interrogation of the parental alleles in terms of CTCF and cohesin binding in unprecedented detail. We examined postnatal day 21 (P21) mouse brain, a time point in development shortly after the growth spurt in neurogenesis that occurs in the first 2 wk of postnatal development (Lyck et al. 2007). In the adult mouse brain, ~56% of cells are neurons and 44% are nonneuronal cells (Fu et al. 2012). Neurons and the principle type of nonneuronal cells, the macroglia, both derive from the neuroepithelium. These data are representative of adult rather than immature brain cell types and are unaffected by long-term aging effects.

Results

We demonstrate that in mouse brain, CTCF and cohesin each bind to ~50,000 sites in the genome, with ~27,000 sites bound by both factors, indicative of CTCF and cohesin acting throughout the genome both in concert as well as independently. Genes are enriched for CTCF binding sites, while intergenic regions are depleted. The binding sites are highly enriched for the canonical consensus binding motif. CTCF binding sites are relatively hypomethylated, both in the CpG and non-CpG sequence context. Parental allele-specific CTCF binding is rare, with most sites at or near imprinted loci. However, a majority but not all DMRs are bound by CTCF (or cohesin), and the binding is not necessarily allele specific. The *Magel2/Peg12* imprinted locus is unique in the genome, comprising a cluster of eight allele-specific CTCF binding sites. Comprehensive expression profiling in mouse brain of genes near allele-specific CTCF binding sites not previously associated with imprinting did not reveal novel imprinted genes. No allele-specific cohesin binding sites of genome-wide significance were found, although at allele-specific CTCF binding sites, there is a trend for cohesin to bind the same allele.

Deep ChIP-seq for CTCF and cohesin to detect parental allele-specific binding

Sites of CTCF and cohesin binding to DNA were determined genome wide in whole P21 mouse brain by chromatin immunoprecipitation (ChIP) using antibodies specific to CTCF and the RAD21 cohesin subunit followed by high-throughput sequencing (ChIP-seq). The mice were the offspring of crosses between C57BL/6 (Bl6) females and *Mus musculus castaneus* (cast) males (B × C), and vice versa (C × B) (Fig. 1A). We generated 235 million and 231 million high-quality and uniquely mapping sequence reads for CTCF and cohesin, respectively (Fig. 1A). The percentage of reads representing clonal duplication was below 6.2% for all samples (Supplemental Fig. S1). Duplicate reads were excluded from further analysis, and regions of CTCF and RAD21 binding were identified using USeq and assigned to either the Bl6 or cast genome based on known SNPs (Fig. 1B; Supplemental Fig. S2; Supplemental Table S1; Keane et al. 2011; Yalcin et al. 2011).

A systematic read mapping bias toward the reference Bl6 genome was observed, consistent with a Bl6 allele read in a polymorphic region being more likely to align for both CTCF and cohesin. However, our use of reciprocal crosses prevented parental allele-specific binding being confounded: There was no overall bias toward either of the parental alleles when the reads generated from both reciprocal crosses were considered together (Fig. 1C,D).

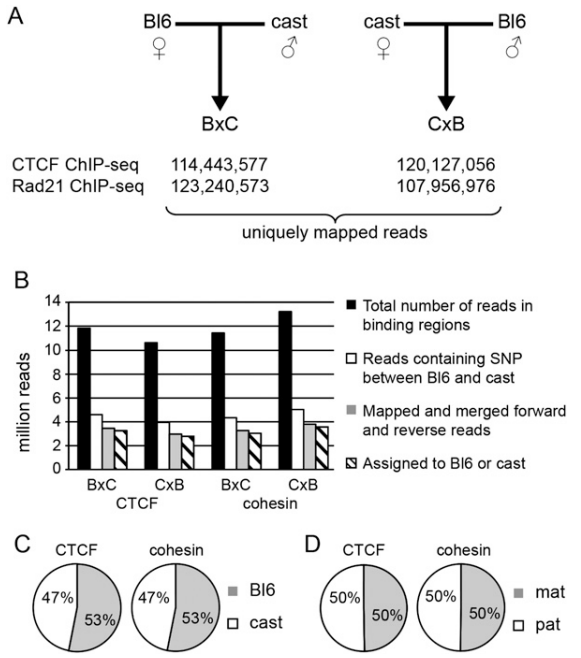


Figure 1. (*A*) ChIP-seq was performed for CTCF and cohesin (RAD21) on P21 brain in B × C and C × B F_1 hybrid animals. The experimental design and number of uniquely mapped reads taken forward for further analysis are shown. (*B*) Regions of CTCF and RAD21 binding were identified using the Useq, and regions identified with a FDR of <13 were considered significant and were tested for parent-of-origin-specific binding. (Black bar) The number of reads for each experiment that fell at, or within ±500 bp of a binding region; (white bar) indicates reads in binding regions that aligned over a SNP between C57BL/6 (Bl6) and *Mus m. castaneus* (cast); (gray bar) the number of reads after the paired reads are considered together and the best-quality read is used to map the read to Bl6 or cast. (Hatched bar) The final number of reads assigned. There was a consistent bias toward the reference sequence (*C*); however, this effect was eliminated after we combined B × C and C × B reads (*D*).

CTCF and cohesin binding in mouse brain

Genome wide, we detected 49,358 CTCF and 52,938 cohesin binding sites with a high degree of statistical confidence. Of these, 27,241 sites were bound by both CTCF and cohesin, accounting for 55.3% of the CTCF and 51.5% of the cohesin binding sites, respectively (Fig. 2). This is consistent with previous studies that show both independent and coordinated roles for these factors (Wendt et al. 2008; Lin et al. 2011).

CTCF binds to regions containing the canonical consensus motif

CTCF binds to a specific DNA sequence motif in ES cells (Chen et al. 2008) and liver (Schmidt et al. 2012). To search for CTCF binding motifs in brain, we applied the MEME de novo motif-finding tool to the sequences of all CTCF binding regions in P21 mouse brain (Fig. 3A). The most significant motif ($P = 2.9 \times 10^{-199}$) is highly similar to the published CTCF binding motif (Chen et al. 2008; Schmidt et al. 2010, 2012). To ensure the consistency of the comparison, we repeated the MEME analysis using identical parameters on CTCF binding regions previously identified using ChIP-seq in ES cells and liver (Chen et al. 2008; Schmidt et al. 2012). Again, the canonical motif was identified as the most significant motif in both ES cells ($P = 7.4 \times 10^{-924}$) and liver ($P = 1.4 \times 10^{-367}$). All three motifs display a high degree of sequence homology, particularly at the core 12-bp sequence at the center of the identified motifs (Fig. 3A).

CTCF binding sites are hypomethylated

The preference of CTCF to bind unmethylated DNA was confirmed by assessing the level of cytosine methylation at CTCF binding sites in brain. Using genome-wide bisulfite-sequencing (BS-seq) data for adult mouse brain (Xie et al. 2012), we compared the overall genome-wide level of methylation at cytosine residues, separately for CpG dinucleotides and non-CpG cytosines, with the portion of the genome corresponding to regions of CTCF binding. We found that methylation at CpG dinucleotides appears to have a greater influence on CTCF binding than non-CpG methylation. Genome wide, 60.8% of CpGs are methylated in the mouse brain, in contrast to 51.9% of CpGs in regions of CTCF binding. Non-CpG methylation also is less frequent in regions of CTCF binding (2.1%) compared with the genome-wide level (2.5%) (Fig. 3B). These differences are statistically significant (χ^2 test, $P < 1 \times 10^{-6}$).

CTCF preferentially binds near genes

We explored the genome-wide location of both CTCF binding regions and parent-of-origin-specific CTCF binding regions using the *cis*-regulatory element annotation (CEAS) tool (Shin et al. 2009). CTCF binding is particularly enriched in regions up to ±3 kb upstream of and downstream from genes, but is depleted in intergenic regions (Fig. 3C). This is consistent with the insulator function of CTCF and, more generally, its involvement in controlling gene expression. When we limited our analysis to parent-of-origin-specific CTCF binding sites, we found the results to be similar. However, intronic regions appeared to be slightly underrepresented and intergenic regions slightly overrepresented relative to the distribution of all CTCF binding sites (Fig. 3C). Given the small number of parent-of-origin-specific CTCF binding sites, these differences are likely due to chance.

Noncanonical CTCF binding sites are tissue specific

We compared the locations of CTCF binding sites in P21 brain with those reported for mouse ES cells and liver (Chen et al. 2008;

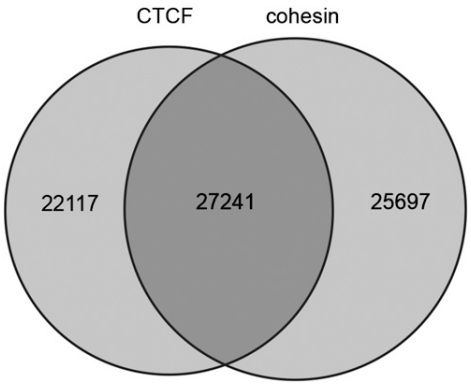
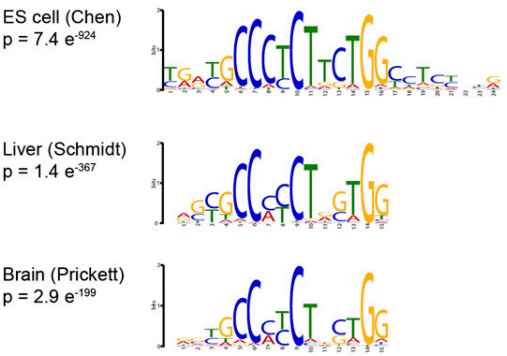


Figure 2. Overlap of the 49,358 CTCF and 52,938 cohesin binding regions in mouse brain. This demonstrates that just over half of CTCF (55%) and cohesin (51%) binding sites are shared, suggesting both independent and combinatorial functions for CTCF and cohesin in the 3-wk mouse brain.

A



B

% cytosine methylation

Cytosine context	Genomewide	Within CTCF binding site	p-value
Within CpG	60.8%	51.9%	$<1x10^{-6}$
Not within CpG	2.5%	2.1%	$<1x10^{-6}$

C

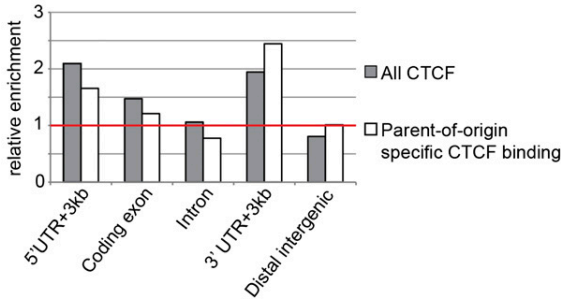


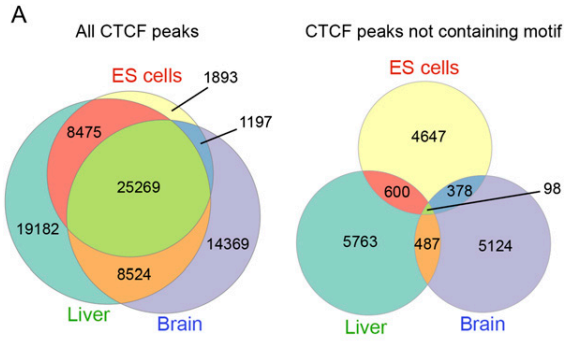
Figure 3. CTCF binding analysis. (*A*) MEME motif finder was executed on the CTCF binding locations identified by ChIP-seq in brain and compared with motifs identified using previously published ES cells and liver binding locations. Each data set found the canonical motif with high degrees of certainty. (*B*) The level of cytosine methylation within CTCF binding sites in the brain was compared with that across the whole genome using data from Xie et al. (2012). In both CpG and non-CpG context cytosine methylation, cytosines within CTCF binding sites are methylated less than those outside CTCF binding sites (χ^2 contingency table tests, $P < 0.001$ for CpG and non-CpG context), confirming that CTCF prefers to bind to unmethylated DNA. (*C*) Genomic locations of CTCF binding are normalized to the proportion of the genome that constitutes each location (represented by the red line). This was considered for all CTCF peaks called with an FDR < 13 and separately for the 116 regions where CTCF binding was seen on one parental allele only (regions identified with a $P < 0.001$). CTCF is significantly enriched at genic regions, but depleted in distal intergenic regions. Parent-of-origin-specific CTCF binding locations are similar but show that binding is depleted in introns but not in intergenic regions.

Schmidt et al. 2012). The incidence of overlap between sites reported in different studies increases with increasing the peak size used for the comparison. Beyond a certain peak size, increases in overlap are mostly due to chance. Therefore, we iteratively increased peak size and compared the incidence of overlap between sites in ES cells, liver, and brain with randomized site locations. Beyond a peak size of 1 kb, increases in the incidence of overlaps were likely due to chance (Supplemental Fig. S3). For a common peak size of 1 kb, 32.0% of all binding sites were shared between ES cells, brain, and liver, suggesting that they are invariant during differentiation and regardless of cell type (Fig. 4A). Only 1893 binding sites were occupied exclusively in ES cells (5.1% of ES-cell binding sites), suggesting that most CTCF binding sites in ES cells represent a ground state that is added to during differentiation, with few binding sites being characteristic of pluripotency per se. In differentiated tissues, 29.1% of brain and 31.2% of liver binding sites are unique to the respective tissue. These analyses were repeated using alternative CTCF ChIP-seq data from ES cells and liver (Shen et al. 2012), producing similar results, even though significantly fewer binding sites were identified in these studies because of limited read depth and quality (Supplemental Fig. S4). We hypothesize that the canonical consensus CTCF binding motif may be at the core of binding sites that are largely invariant with respect to cell type, concordant with other findings (Essien et al. 2009). We restricted the above overlap analysis to CTCF binding sites that lack the canonical binding motif. There was a large reduction in the number of binding sites shared between tissue types (Fig. 4A), with most binding sites now being tissue specific: 84.2% of binding sites in brain that lack the canonical motif were brain specific, and similarly for ES cells (81.2%) and liver (82.9%) (Fig. 4B). These results suggest that CTCF binding to tissue-specific sites may involve other consensus motifs recognized by cofactors or tissue-specific conformations of the 11 zinc finger domains of CTCF itself.

Parent-of-origin-specific CTCF and/or cohesin binding is limited to specific DMRs

We systematically investigated the binding of CTCF and cohesin at or near 22 known well-characterized mouse gDMRs (Table 1) associated with imprinted gene expression (most of which are classified as ICRs). Of the 22 gDMRs (Table 1), 19 have a CTCF and/or cohesin binding site in brain within 2.5 kb. Of these sites, 12 are bound by both CTCF and cohesin, three by CTCF alone and four by cohesin alone. gDMRs with both CTCF and cohesin binding

A



B

% unique peaks

	ES cell	Liver	Brain
All CTCF peaks	5.1%	31.2%	29.1%
CTCF peaks not containing motif	81.2%	82.9%	84.2%

Figure 4. (*A*) Proportional Venn diagrams comparing coincidence of CTCF binding sites between ES cells, liver, and brain demonstrate significant overlap of CTCF binding in these tissues, Coincident binding was also considered after the removal of binding regions containing the consensus CTCF motif; overlap of CTCF binding in the absence of the consensus motif was much lower than when all binding sites were considered. (*B*) The percentages of shared peaks for each tissue type for all peaks and for nonmotif peaks.

Table 1. Gene names for imprinted regions are listed with corresponding gDMR positions

gDMR information (Wamidex)			CTCF and cohesin binding by ChIP-seq						Previous evidence
			CTCF			Cohesin			
gDMR name	gDMR position (Wamidex)	Methylated allele	CTCF binding?	Binding allele	Allele specific *P*-value	Cohesin binding?	Binding allele	Allele binding *P*-value	Reference
CTCF and cohesin precisely colocalized at gDMR									
Grb10	chr11: 11,925,485–11,925,790	Maternal	Yes	N/A	N/A	Yes	N/A	N/A	Hikichi et al. 2003
H19/Igf2	chr7: 149,766,168–149,768,424	Paternal	Yes	Maternal	1.11×10^{-74}	Yes	Maternal	1.08×10^{-4}	For example, Hark et al. 2000
Inpp5f_v2	chr7: 135,831,788–135,832,156	Maternal	Yes	Bi-allelic	N/A	Yes	Bi-allelic	N/A	
Mcts2	chr2: 152,512,491–152,513,011	Maternal	Yes	N/A	N/A	Yes	N/A	N/A	
Mest	chr6: 30,686,709–30,687,273	Maternal	Yes	Paternal	9.00×10^{-15}	Yes	Paternal	0.0414	
Nnat	chr2: 157,385,786–157,387,398	Maternal	Yes	N/A	N/A	Yes	N/A	N/A	
Peg13	chr15: 72,636,765–72,642,079	Maternal	Yes	Paternal	1.35×10^{-57}	Yes	Paternal	5.50×10^{-3}	Singh et al. 2011
Plagl1	chr10: 12,810,276–12,810,604	Maternal	Yes	Bi-allelic	N/A	Yes	Bi-allelic	N/A	
CTCF and cohesin not precisely colocalized at gDMR									
Cdh15	chr8: 125,387,861–125,390,344	Maternal	Yes	Paternal	0.0463	Yes	N/A	N/A	
Nespos	chr2: 174,121,208–174,126,482	Maternal	Yes	N/A	N/A	Yes	N/A	N/A	
Zrsr1	chr11: 22,871,842–22,872,319	Maternal	Yes	N/A	N/A	Yes	Bi-allelic	N/A	
Zim2 (Peg3)	chr7: 6,680,287–6,684,827	Maternal	Yes	Paternal	1.16×10^{-30}	Yes	Paternal	0.049	
CTCF binding only									
Peg10	chr6: 4,697,209–4,697,507	Maternal	Yes	N/A	N/A	No	N/A	N/A	
Meg3/Dlk1	chr12: 110,761,563–110,768,989	Paternal	Yes	N/A	N/A	No	N/A	N/A	
Impact	chr18: 13,130,706–13,132,250	Maternal	Yes	N/A	N/A	No	N/A	N/A	
Cohesin binding only									
Igf2r-air	chr17: 12,934,163–12,935,573	Maternal	No	N/A	N/A	Yes	N/A	N/A	
Gnas-exon1A	chr2: 174,153,279–174,153,502	Maternal	No	N/A	N/A	Yes	N/A	N/A	
Kcnq1ot1	chr7: 150,481,060–150,481,397	Maternal	No	N/A	N/A	Yes	N/A	N/A	Fitzpatrick et al. 2007
Snurf/Snrpn	chr7: 67,149,878–67,150,301	Maternal	No	N/A	N/A	Yes	N/A	N/A	
No binding									
Nap1l5	chr6: 58,856,690–58,857,056	Maternal	No	N/A	N/A	No	N/A	N/A	
Rasgrf1	chr9: 89,774,406–89,774,691	Paternal	No	N/A	N/A	No	N/A	N/A	Yoon et al. 2005
Slc38a4	chr15: 96,885,270–96,886,284	Maternal	No	N/A	N/A	No	N/A	N/A	

The methylated allele is shown and CTCF binding is indicated according to the criteria used in this study. The parental allele specificity is shown, but for some regions, these data did not reach significance, or "no data" was registered, which was due either to the absence of an SNP or insufficient sequence reads over the SNP. *P*-values are given for the allele-specific binding, where $P < 0.05$. (Gray shading) Genome-wide significant parent-of-origin-specific binding. References to previous binding are indicated where known.

sites within 2 kb formed two categories: those where CTCF and cohesin colocalized precisely at the gDMR (eight regions), and a further four regions where binding occurred near but not over the gDMR and CTCF and cohesin each bound distinct sites (Table 1). Where the two factors are precisely colocalized on the DNA, cohesin binding is probably linked mechanistically to CTCF, while this is less likely at gDMRs where binding sites do not coincide.

For gDMRs where genome-wide significant ($P < 1 \times 10^{-6}$) parent-of-origin-specific binding events for CTCF (Table 1) were detected (*H19/Igf2*, *Mest*, *Peg13*, and *Zim2* [*Peg3*]), binding occurred as expected on the unmethylated allele (Table 1). The *Mest* and *Zim2* gDMRs were not previously known to bind CTCF in a parent-of-origin-specific manner. In addition, the 95% confidence intervals (Supplemental Fig. S5; Supplemental Table S2) for parent-of-origin-specific binding showed a trend toward preferential binding of CTCF to the unmethylated alleles of the *Grb10*, *Mcts2*, *Cdh15*, *Nespos*, *Zrsr1*, *Peg10*, and *Meg3/Dlk1* gDMRs. We considered CTCF binding to be completely biallelic if the 95% confidence interval for the maternal-over-paternal read ratio was between 0.35 and 0.65 and spanned 0.5. This was the case for the *Inpp5f_v2* and *Plagl1* gDMRs. At *H19/Igf2* and *Peg13*, parent-of-origin-specific binding of cohesin was detected but was not genome-wide significant after Bonferroni multiple testing correction (Table 1). The overall pattern of the 95% confidence intervals for the ratio of maternal-to-paternal reads for CTCF and cohesin suggests that in comparison to CTCF, cohesin binding is less biased toward the unmethylated parental allele (Supplemental Fig. S5; Supplemental Table S2). This is consistent with increased recruitment of cohesin to sites bound by CTCF.

Parent-of-origin-specific CTCF and cohesin binding at gDMRs could only be tested where a Bl6-cast SNP is within the bound region (Table 1). In addition, CTCF and cohesin peaks did not always overlap perfectly so that for some gDMRs, a SNP was informative for one factor but not the other. Another limitation for the detection of parent-of-origin-specific binding arose when a SNP was located at the periphery of the respectively bound region where fewer reads align and the statistical power of the binomial test was diminished. For CTCF, these limitations applied in particular to the *Grb10*, *Mcts2*, *Nnat*, *Nespos*, *Zrsr1*, *Impact*, and *Peg10* gDMRs. For cohesin, the above limitations applied to half of the cohesin-bound gDMRs (Supplemental Table S2). The results for CTCF support the notion that it plays a central role in imprinting control at several loci. This is in contrast to cohesin, in particular, four gDMRs (*Gnas-exon1A*, *Igf2r-air*, *Kcnq1ot1*, and *Snurf/Snrpn*) were bound by cohesin but not by CTCF; here binding was not parental allele specific. Cohesin binding independently of CTCF is not unprecedented (Schmidt et al. 2012), and there is evidence that it is more generally involved in transcriptional activation (Kagey et al. 2010).

Genome-wide, parental allele-specific binding of CTCF and cohesin is rare and mostly restricted to imprinted loci

CTCF binding efficiency to methylated DNA is reduced compared with unmethylated DNA (Mukhopadhyay et al. 2004) explaining parental allele-specific binding at the *H19/Igf2* and other gDMRs. If CTCF and cohesin are exerting a key regulatory role at several imprinted loci, then genome wide, other occurrences of parental allele-specific CTCF and/or cohesin binding may identify novel DMRs and imprinted genes. Four known ICRs—*H19/Igf2*, *Peg13*, *Zim2* (*Peg3*), and *Mest*—met the genome-wide significance threshold for parental allele-specific CTCF binding providing proof of principle. Only an additional 17 regions reached genome-wide significance (Fig. 5; Supplemental Fig. S3; Table 2). Eight of these sites clustered in a 250-kb region on chromosome 7 at the *Peg12/Magel2* imprinted domain (Fig. 6). A further four sites were within 6 Mb of other known imprinted regions. Two more are 30 kb apart on chromosome 14 (Fig. 5). Many chromosomes were devoid of parental allele-specific CTCF binding, and no cohesin binding regions were detected at genome-wide significance (Supplemental Table S3). Of all 21 genome-wide significant parental allele-specific CTCF binding sites, six were on the maternal and 15 on the paternal allele. We tested all gene transcripts at or near these sites not previously reported as imprinted for parental allele-specific expression in mouse brain. Many showed a complex organization of transcripts (Supplemental Fig. S6), but none were imprinted (Supplemental Table S4).

Eight sites of parental allele-specific CTCF binding at the *Peg12/Magel2* imprinted domain (Fig. 6) bound CTCF on the paternal allele, indicating maternal methylation. We assayed methylation of the CpG island at the promoter of *Magel2*, which is in close proximity to two CTCF sites and maternally methylated (Supplemental Fig. S7). This is confirmation of the parental allele-specific methylation of the region recently reported (Xie et al. 2012). The *Magel2* DMR is likely somatic and established post-fertilization that is supported by genome-wide methylation data in oocytes (Smallwood et al. 2011). In addition, *Dnmt3L*$^{-/+}$ 8.5 days postcoitum (dpc) embryos are unchanged at the *Magel2* promoter relative to wild type and are unmethylated (Proudhon et al. 2012). This suggests that the maternal allele-specific methylation at the *Magel2* promoter, and presumably the other sites of paternal allele-specific CTCF binding in the domain, is established post-implantation and/or is brain specific. The regulation of the imprinted domain comprising *Ndn*, *Magel2*, *Mkrn3*, and *Peg12* deserves further study since the human orthologs of *Ndn*, *Magel2*, and *Mkrn3* are in the region associated with Prader-Willi syndrome (Lee and Wevrick 2000), with patients displaying a notable range of neurological symptoms. Given this extensive investigation of parental allele-specific CTCF binding, we predict that there are few additional DMRs in the adult mouse brain bound by CTCF.

Validation of CTCF and cohesin binding at specific loci

Quantitative assays for *H19/Igf2*, *Peg10*, *Nap1l5*, *Nnat*, and *Grb10* DMR validated that the ChIP-seq data (Supplemental Fig. S8A,B)

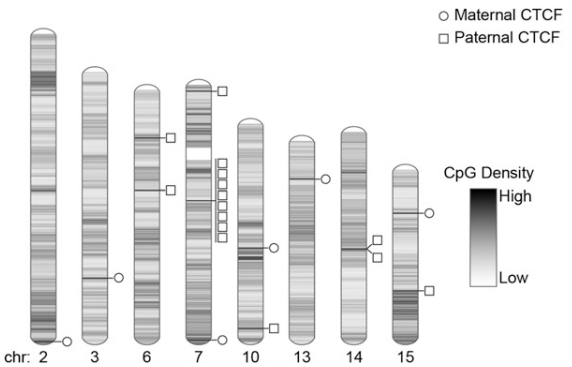


Figure 5. Chromosomal location of genome-wide significant parent-of-origin-specific CTCF binding regions. Where CTCF is bound on the maternally inherited allele, this is illustrated with a circle; where CTCF is bound on the paternally inherited allele, this is illustrated with a square. Only chromosomes where parent-of-origin-specific binding was seen are shown. CpG density is indicated.

Table 2. 21 regions were identified where CTCF binds to one allele in a parent-of-origin-specific manner

Location (mm9 reference)	Binding allele	*P*-value	Ratio of binding:nonbinding	Nearest genes	Notes
chr7: 149,764,416–149,768,874	Maternal	1.11×10^{-74}	10.3	*H19*, *Igf2*	Known imprinted gene region
chr15: 72,638,890–72,641,957	Paternal	1.35×10^{-57}	12.6	*Peg13*, *Trappc9*	Known imprinted gene region
chr7: 6,678,325–6,681,689	Paternal	1.17×10^{-30}	5.5	*Zim2* (*Peg3*)	Known imprinted gene region
chr6: 30,686,300–30,688,046	Paternal	9.10×10^{-15}	9.4	*Mest*	Known imprinted gene region
chr7: 69,543,049–39,547,037	Paternal	1.35×10^{-12}	3.4	*Magel2*	Known imprinted gene region
chr7: 69,580,613–69,582,990	Paternal	2.05×10^{-10}	5.6	*Magel2*	Known imprinted gene region
chr7: 69,323,407–69,325,218	Paternal	3.91×10^{-10}	6.6	*Magel2*	Known imprinted gene region
chr10: 74,395,653–74,404,537	Maternal	1.28×10^{-9}	1.5	*Rtdr1*, *Gnaz*	No known imprinted genes within 20 Mb
chr7: 69,526,343–69,528,366	Paternal	3.10×10^{-9}	4.1	*Magel2*	Known imprinted gene region
chr7: 69,372,124–69,373,922	Paternal	3.26×10^{-9}	8.0	*Ndn*, *Magel2*	Known imprinted gene region
chr2: 180,079,574–180,091,367	Maternal	5.84×10^{-9}	1.5	*Gata5*, *Gm14318*	6 Mb from *Gnas* locus
chr7: 69,608,918–69,610,897	Paternal	6.90×10^{-9}	5.6	*Peg12*, *Mkrn3*	Known imprinted gene regions
chr14: 69,941,084–69,946,555	Paternal	1.68×10^{-8}	1.5	*Gm16677*, *Entpd4*, *Loxl2*	5 Mb from *Htr2a* imprinted locus
chr7: 69,353,580–69,355,185	Paternal	2.15×10^{-8}	5.1	*Ndn*, *Magel2*	Known imprinted gene regions
chr7: 69,519,941–69,521,489	Paternal	3.38×10^{-8}	5.3	*Magel2*	Known imprinted gene regions
chr14: 69,994,239–70,003,685	Paternal	3.98×10^{-8}	1.5	*Entpd4*, *AK086749*, *Loxl2*	5 Mb from *Htr2a* imprinted locus
chr10: 120,737,183–120,739,873	Paternal	4.75×10^{-8}	2.7	*Tbc1d30*	No known imprinted genes within 20 Mb
chr3: 121,236,161–121,244,419	Maternal	6.85×10^{-8}	1.7	*A530020G20Rik*, *Slc44a3*	No known imprinted genes on chromosome 3
chr15: 27,817,069–27,819,622	Maternal	6.95×10^{-8}	2.3	*Trio*	No known imprinted genes within 20 Mb
chr13: 25,098,042–25,100,314	Maternal	9.98×10^{-8}	2.1	*Mrs2*, *Gpld1*	No known imprinted genes within 20 Mb
chr6: 60,631,333–60,634,328	Paternal	2.47×10^{-7}	2.2	*Snca*	1.6 Mb from *Herc3*

After correction for multiple testing, regions are ranked in order of statistical significance (*P*-value). Twelve regions are associated with known imprinted genes, of which eight are associated with the *Peg12/Magel2* imprinted locus. Four further regions occur within close proximity of an imprinted locus. All novel candidates were tested for imprinting (Supplemental Table S4).

results were in agreement (Table 1) with the exception of CTCF binding at *Nnat* and cohesin binding at *Peg10*. Both are borderline cases. Using qPCR to detect CTCF binding at the *Nnat* DMR resulted in $P = 0.08$, just above our cutoff for binding, At *Peg10*, RAD21 binding was detected by qPCR, but no peak was identified by ChIP-seq. When the stringency of the ChIP-seq peak detection is relaxed, two RAD21 binding regions ~1 kb either side of the qPCR regions are detected (Supplemental Fig. S8C).

Validation of parental allele-specific binding

To validate allele-specific binding we pyrosequenced ChIP'd mouse brain from reciprocal crosses. We selected three representative DMRs, based on the ChIP-seq results: *Inpp5f_v2*, where biallelic CTCF binding was detected; *Mest*, where we detected paternal allele-specific binding; and *Peg10*, where the CTCF binding site did not meet the significance threshold for allele-specific binding but where the 95% confidence interval was suggestive of CTCF binding on the paternal allele. These results agreed with our ChIP-seq data (Supplemental Fig. S9): *Inpp5f_v2* does not deviate from the expected 50:50 allelic ratio ($P = 0.3214$), *Mest* shows paternal binding ($P = 0.0017$), and *Peg10* shows a bias toward enrichment of the paternal allele ($P = 0.0813$).

CTCF and cohesin binding at somatic DMRs

A set of 23 known somatic and novel putative somatic DMR coordinates has recently been defined by whole-genome bisulfite sequencing (BS-seq) in mouse brain (Xie et al. 2012). We evaluated CTCF and cohesin binding in the somatic DMRs identified in this study (Supplemental Table S5). We found 13 instances of CTCF binding, two of which were parental allele specific ($P < 1 \times 10^{-6}$) and 14 instances of cohesin binding, none of which were parental allele specific. All parental allele-specific binding involved the

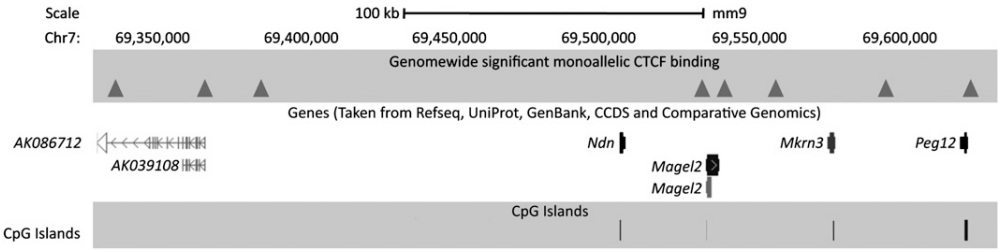


Figure 6. Multiple parent-of-origin-specific CTCF binding sites are observed on the paternal allele at the *Magel2/Peg12* locus. (Triangles) Paternally bound CTCF binding sites. Genes and CpG islands are indicated. This region represents a unique example in the mouse genome of CTCF bound only on the paternal allele at eight regions in close proximity. This figure was adapted from the UCSC Genome Browser.

unmethylated allele. Overall, the results for somatic DMRs (Supplemental Table S6) are in close agreement with the results for gDMRs (Table 1) so that the origin of a DMR, germline versus somatic, is not a determinant of CTCF and/or cohesin involvement in the regulation of imprinting.

Discussion

CTCF and cohesin act synergistically and independently in mouse brain and show tissue-specific distributions compared with undifferentiated cells

Several studies have examined the colocalization of CTCF with cohesin (Chen et al. 2008; Parelho et al. 2008; Rubio et al. 2008; Kagey et al. 2010; Schmidt et al. 2012), and CTCF physically associates with cohesin via the Stag1 (Scc3/SA1) subunit in human cells (Wendt et al. 2008). Here 55% of CTCF binding sites overlap with cohesin binding, and the remaining sites binding independently (Fig. 2), indicating that CTCF fulfils a role independent from as well as in combination with cohesin in brain. This supports the idea that different functions may be a result of the context of CTCF binding (Gaszner and Felsenfeld 2006), and it is possible that coordinate binding of cohesin may influence CTCF. Cohesin is involved in tissue-specific transcriptional control (Faure et al. 2012) and associated with the Mediator complex, which has a role in transcriptional activation (Taatjes 2012). Studies have shown a link between cohesin, the Mediator complex, transcription, and chromatin looping (Kagey et al. 2010). We report that 51% of cohesin sites in brain are not coincident with CTCF, consistent with CTCF not being required for the loading of cohesin onto DNA (Rubio et al. 2008).

This comparison of CTCF binding in ES cells, liver, and brain reveals more unique CTCF binding sites in differentiated cells than in ES cells, suggesting tissue-specific CTCF binding in the specification and/or maintenance of differentiated tissue. We observe a significant overlap in CTCF binding between tissues (Fig. 4A), consistent with studies reporting highly conserved CTCF binding between cell types (Kim et al. 2007b).

CTCF binding

In brain, CTCF binds to unmethylated regions of DNA, usually to the canonical CTCF motif (Chen et al. 2008; Schmidt et al. 2010, 2012). CTCF binding outside this motif is much more tissue specific, and there is little overlap between tissues (Fig. 4A). This is consistent with evidence from CTCF knockout mouse studies, which exhibit embryonic lethality prior to implantation (Splinter et al. 2006). We found that the canonical consensus binding motif is most frequent at CTCF binding sites shared between ES cells, brain, and liver; thus, it is associated with invariant binding during differentiation. CTCF binding appears overrepresented just up or downstream from gene bodies with a paucity in distal intergenic regions, unsurprising given the known role of CTCF in gene expression (Bell et al. 1999; Cuddapah et al. 2009). CTCF mediates long-range chromosomal interactions genome wide in *cis* and in *trans* in ES cells (Handoko et al. 2011), and we provide additional evidence genome wide that CTCF is an insulator at or near gene coding regions by binding to noncoding DNA.

Cytosine methylation at CTCF binding regions

Cytosine methylation in both a CpG and non-CpG context is reduced in regions of CTCF binding compared with the level observed genome wide, consistent with published data (Mukhopadhyay et al. 2004). Interestingly the canonical motif lacks CpG dinucleotides, suggesting that methylation of DNA in the motif does not preclude CTCF binding, but surrounding methylation is important. The canonical motif may not function alone, but in concert with another region of DNA ~20 bp downstream, suggesting that CTCF interaction with DNA is not limited to the 20-mer motifs (Schmidt et al. 2012).

Parent-of-origin-specific CTCF and cohesin binding

CTCF and cohesin bind at numerous imprinting control regions and other DMRs as previously detected, but not systematically tested. The presence of CTCF and cohesin together at 12 imprinting associated gDMRs in brain (Table 1) is consistent with a regulatory role for these proteins at imprinted loci. Studies using 3C and 4C have shown that several imprinted domains are physically clustered (Sandhu et al. 2009), in part because CTCF (Botta et al. 2010) and cohesin (Murrell et al. 2004; Nativio et al. 2009) form loops that contribute to three-dimensional (3D) nuclear architecture (Phillips and Corces 2009). The CTCF and cohesin binding (Table 1) supports the idea of three types of imprinting mechanisms: CTCF dependent, CTCF/cohesin mediated, and CTCF/cohesin independent.

CTCF and cohesin bind at somatic DMRs, suggesting a role for them here. Parental allele-specific binding of CTCF together with cohesin regulates allele-specific expression in somatic cells (Lin et al. 2011), while cohesin binding alone may be involved in the transcriptional regulation of imprinted gene expression generally. CTCF and cohesin are likely to have distinct functions in different cell types at a subset of targets (Lin et al. 2011), and findings in mouse brain support the idea that these proteins play a role in imprinting at some loci and at others they act more generally.

These data provide a resource for interrogating the roles of CTCF and cohesin and point to a role at more imprinted loci than was previously appreciated, although further functional studies would be needed to confirm this. Three gDMRs do not bind either factor, illustrating the heterogeneous nature of gDMRs as a group of regulatory regions. For example, the four imprinted retrogene/host gene pairs *Mcts2/H13*, *Nap1l5/Herc3*, *Inpp5f/Inpp5f_v2*, and *Zrsr1/Commd1* share several sequence-based and genomic context–related features (Wood et al. 2007). Since within this group, *Mcts2* binds both CTCF and cohesin together, *Nap1l5* binds neither CTCF nor cohesin, *Inpp5f_v2* binds CTCF on both parental alleles equally, and *Zrsr1* binds both CTCF and cohesin, but they do not colocalize, this suggests no consistent mechanism for imprinting control despite the other shared features.

CTCF binding profiles vary between different tissues. We show that many CTCF binding sites are shared between ES cells and differentiated tissues and that this type of invariant CTCF binding is associated with the canonical CTCF motif. CTCF binding in the absence of the canonical motif is associated with tissue-specific CTCF binding.

Methods

Chromatin immunoprecipitation

Chromatin from whole tissue was isolated, sonicated, and immunoprecipitated for ChIP-seq library preparation according to Supplemental Methods 1.

Next-generation sequencing

Library preparation

DNA enriched through ChIP was quantified using the Qubit (Invitrogen) and Quant-iT dsDNA high-sensitivity assay kit (Invitrogen:Q32854) and was sized using the Agilent Bioanalyzer with a High Sensitivity DNA Bioanalyzer kit (5067-4626). DNA was fragmented to a size appropriate for the library preparation step using the Covaris S220, samples were sheared over two cycles: 5% duty cycle, 3 intensity, 200 cycles per burst, and time of 65 sec. DNA from ChIPs performed on chromatin extracted from two mice was pooled.

ChIP-seq libraries were prepared using the Illumina ChIP-seq library preparation kit (IP-102-1001) and the NEBNext ChIP-seq library preparation kit (E6240). Libraries were sized and quantified using an Agilent Bioanalyzer and a High Sensitivity Kit (5067-4626).

ChIP–seq data analysis

Sequence reads were aligned to the mouse reference genome (mm9) using Novoalign (v. 2.01.13; http://www.novocraft.com/). USeq (Nix et al. 2008) was used to identify mean peak shift separately for CTCF and cohesin reads using only the first of each pair-end matched read. Peaks were identified using peak shifts and window sizes of 138 bp and 144 bp for CTCF and cohesin, respectively, and a False Discovery Rate (FDR) of 95% (Supplemental Table S2A). A subset of peaks was obtained to a false discovery ratio of 5% (Phred-scaled FDR 13) expanded by 500 bp upstream and downstream and overlapping peaks merged prior to further analysis. Refer to Supplemental Table S2A for the number of raw reads that pass a quality control map in a CTCF or cohesin binding region.

Parental allele–specific binding analysis

Parental allele-specific binding was assessed by binomial testing, using a custom bioinformatics pipeline. For performance reasons, only reads of interest, which overlapped the previously identified CTCF or RAD21 binding sites and a SNP between the two parental strains, were extracted from the SAM files and used for subsequent analysis.

Individual reads were assigned to one of the parental alleles using a custom Perl script, using the SAMtools Perl library. Each read was mapped as either derived from the reference sequence (Bl6) or from the cast allele on the basis of a SNP between the parental strains. If more than one SNP was present, the SNP with the best quality of read sequence was used. Reads were only considered for subsequent analysis if the Phred-scaled alignment mapping quality exceeded 50 and the base call quality at the SNP used for mapping of the read exceeded 20.

Paired reads were mapped to parental strains separately and merged. Because paired reads are not independent data points, when they were in disagreement (<1%) the read pair was assigned on the basis of the best SNP in either of the two reads.

Assigned reads were converted to maternally or paternally derived, and data from both B × C and C × B reciprocal crosses were merged for the CTCF and RAD21 data sets independently. Counts of maternal and paternal reads were obtained on a per-region basis using MySQL. Binding regions were only tested for parent-of-origin-specific expression if three or more reads could be mapped.

Parental allele-specific binding was assessed using a two-sided binomial test (implemented in R) of the maternal-versus-paternal allelic read counts. Regions were sorted by *P*-value score using MySQL. The genome-wide significance of *P*-values was assessed by means of Bonferroni correction. UCSC BED tracks were prepared at different cutoffs with maternal/paternal annotation.

Peak intersections

All subsequent bioinformatic analyses were performed on expanded regions unless otherwise specified. CTCF peak intersections between ES cell, brain, and liver data were performed using an optimized peak size of 1 kb for all data sets. ES cell data were converted to mm9 using the UCSC liftOver tool. Peak overlap counts were obtained used the BEDTools intersectBed command. For each intersection, counts of the intersecting peaks were calculated in both possible ways, and the peaks count reported was the mean of the two measurements. For intersections of more than two data sets, only one of the possible configurations of intersections was examined; the same configuration was used for all analyses.

Identification of non–motif–containing peaks

Peaks that did not contain the CTCF motif were identified using the FIMO tool from the MEME suite (Grant et al. 2011). Peak sequences were obtained using the UCSC Genome Browser table tool in FASTA format with repeat sequences masked. The CTCF motif identified from the brain data set was used throughout, and the threshold for detection was set to 10^{-3}. Custom UNIX shell scripts were used to extract the coordinates of the peaks from the FIMO output.

Motif finding

Motif finding was performed using MEME (Bailey et al. 2009) on binding regions using default MEME parameters. For the brain data, the best subwindow coordinates were used.

Genomic distribution of peaks

The CEAS tool (Shin et al. 2009) was used to assess the genomic distribution of unexpanded CTCF binding regions, and unexpanded parent-of-origin-specific CTCF binding sites were detected with a $P < 0.001$. Relative abundance was normalized to the proportion of the genome represented by each genomic region. For the CEAS analysis, parent-of-origin-specific CTCF expanded regions were assigned back to their original constituent unexpanded peaks using bedmap (Neph et al. 2012).

Quantitative PCR validation of CTCF and cohesin ChIP–seq

These assays are detailed in Supplemental Methods 1.

Validation of parent–of–origin–specific binding using pyrosequencing

Chromatin was extracted from four biological replicates, two B × C and two C × B P21 brains. Pyrosequencing validated CTCF binding at three regions (Supplemental Table S6). ChIP was performed as for ChIP-seq. Maternal-to-paternal proportions were assigned using SNPs between Bl6 and cast. Allelic proportions were normalized to input DNA, which represents a 50:50 ratio of maternal-to-paternal reads. Using the normalized maternal proportion, a two-sided *t*-test against a 0.5 null proportion was performed.

Testing for imprinted expression

Transcripts were tested for allele-specific expression using PCR followed by Sanger sequencing (using the primers in Supplemental Table S7).

Bisulfite mutagenesis

Genomic DNA from B × C and C × B intercross mouse brain tissue was converted using the Zymo EZ DNA Methylation-Direct Kit (D5020). Amplified regions of interest were ligated into pGEM-T Easy (Promega:A1360), transformed into competent *Escherichia coli*, and sequenced. Primers were designed with MethPrimer (For:GTGTTTGTTGAGAGTTGTTGAGAGA; Rev: ACCAAACAACC ATAAAAACCTACAA) (Li et al. 2002).

Data access

Primary sequencing data have been deposited in the NCBI Gene Expression Omnibus (GEO; http://www.ncbi.nlm.nih.gov/geo/) under accession number GSE35140.

Acknowledgments

We thank Dr. Michael Cowley and Heba Saadeh for their critical reading of this manuscript and Dr. Sabrina Böhm for competent cells. We thank Dr. Deborah Bourc'his for providing methylation data from *Dnmt3L*$^{+/-}$ embryos. We thank Seth Seegobin and Dr. Jennifer Mollon for discussions about the value of statistical methodologies for qPCR and pyrosequencing analyses, for discussions on confidence intervals and *t*-tests, and Dr. Michael Weale for discussions about genic enrichment measurements for binding proteins. We acknowledge funding by the Wellcome Trust (Grant number 084358/Z/07/Z to R.J.O.), the Philip Harris prize studentship fund (R.B.M.), the British Heart Foundation studentship FS/08/051/25748 and latterly the MRC (to R.J.O. supporting A.R.P.), the BBSRC Doctoral Training Grant (S.H.), the RCUK fellowship program (R.S.), and King's College London for a partial PhD studentship (N.B.). We acknowledge financial support from the Department of Health via the National Institute for Health Research (NIHR) comprehensive Biomedical Research Centre award to Guy's & St Thomas' NHS Foundation Trust in partnership with King's College London and King's College Hospital NHS Foundation Trust. Their support is through a partial PhD studentship (N.B.), access to the high-performance computing cluster through Don Lokuadassuriyage (Systems Administrator), and services provided by the genomics core facility (from Mudassa Mirza and Dr. Efterpi Papouli) including high-throughput sequencing and access to equipment. We also thank Dr. Matt Arno and Dr. Estibaliz Aldecoa-Otalora Astarloa from the KCL genomic facility, and Dr. Charles Mein from the Barts and the London Genome Centre for pyrosequencing support.

Author contributions: A.R.P. carried out the ChIP-seq experimental work, made the majority of the data acquisition, performed some of the analyses, and drafted part of the paper. N.B. performed the majority of the bioinformatic analyses, made the tables and figures, and revised some of the paper. R.B.M. contributed to the conception and initial experimental design and to initial technique development and some bioinformatics and drafted part of the paper. S.H. performed the bisulfite analysis. S.M.M. performed the allele-specific assays. R.S. contributed to the conception and design of the experiments and bioinformatic analyses and wrote part of the paper. R.J.O. contributed to the conception and design of the experiments and wrote part of the paper. All authors contributed to the final version of the paper.

References

Bailey TL, Bodén M, Buske FA, Frith M, Grant CE, Clementi L, Ren J, Li WW, Noble WS. 2009. MEME SUITE: Tools for motif discovery and searching. *Nucleic Acids Res* **37:** W202–W208.

Bartolomei M. 2009. Genomic imprinting: Employing and avoiding epigenetic processes. *Genes Dev* **23:** 2124–2133.

Bell A, Felsenfeld G. 2000. Methylation of a CTCF-dependent boundary controls imprinted expression of the Igf2 gene. *Nature* **405:** 482–485.

Bell A, West A, Felsenfeld G. 1999. The protein CTCF is required for the enhancer blocking activity of vertebrate insulators. *Cell* **98:** 387–396.

Botta M, Haider S, Leung IX, Lio P, Mozziconacci J. 2010. Intra- and inter-chromosomal interactions correlate with CTCF binding genome wide. *Mol Syst Biol* **6:** 426.

Cassidy S, Dykens E, Williams C. 2000. Prader-Willi and Angelman syndromes: Sister imprinted disorders. *Am J Med Genet* **97:** 136–146.

Chen X, Xu H, Yuan P, Fang F, Huss M, Vega V, Wong E, Orlov Y, Zhang W, Jiang J, et al. 2008. Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. *Cell* **133:** 1106–1117.

Cuddapah S, Jothi R, Schones DE, Roh TY, Cui K, Zhao K. 2009. Global analysis of the insulator binding protein CTCF in chromatin barrier regions reveals demarcation of active and repressive domains. *Genome Res* **19:** 24–32.

da Rocha S, Edwards C, Ito M, Ogata T, Ferguson-Smith A. 2008. Genomic imprinting at the mammalian Dlk1–Dio3 domain. *Trends Genet* **24:** 306–316.

Davies W, Isles A, Humby T, Wilkinson L. 2007. What are imprinted genes doing in the brain? *Epigenetics* **2:** 201–206.

DeVeale B, van der Kooy D, Babak T. 2012. Critical evaluation of imprinted gene expression by RNA-seq: A new perspective. *PLoS Genet* **8:** e1002600.

Edwards C, Ferguson-Smith A. 2007. Mechanisms regulating imprinted genes in clusters. *Curr Opin Cell Biol* **19:** 281–289.

The ENCODE Project Consortium. 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489:** 57–74.

Essien K, Vigneau S, Apreleva S, Singh LN, Bartolomei MS, Hannenhalli S. 2009. CTCF binding site classes exhibit distinct evolutionary, genomic, epigenomic and transcriptomic features. *Genome Biol* **10:** R131.

Faure AJ, Schmidt D, Watt S, Schwalie PC, Wilson MD, Xu H, Ramsay RG, Odom DT, Flicek P. 2012. Cohesin regulates tissue-specific expression by stabilizing highly occupied *cis*-regulatory modules. *Genome Res* **22:** 2163–2175.

Fedoriw A, Stein P, Svoboda P, Schultz R, Bartolomei M. 2004. Transgenic RNAi reveals essential function for CTCF in H19 gene imprinting. *Science* **303:** 238–240.

Filippova GN, Fagerlie S, Klenova EM, Myers C, Dehner Y, Goodwin G, Neiman PE, Collins SJ, Lobanenkov VV. 1996. An exceptionally conserved transcriptional repressor, CTCF, employs different combinations of zinc fingers to bind diverged promoter sequences of avian and mammalian c-*myc* oncogenes. *Mol Cell Biol* **16:** 2802–2813.

Fitzpatrick G, Pugacheva E, Shin J-Y, Abdullaev Z, Yang Y, Khatod K, Lobanenkov V, Higgins M. 2007. Allele-specific binding of CTCF to the multipartite imprinting control region KvDMR1. *Mol Cell Biol* **27:** 2636–2647.

Frost J, Monk D, Moschidou D, Guillot PV, Stanier P, Minger SL, Fisk NM, Moore HD, Moore GE. 2011. The effects of culture on genomic imprinting profiles in human embryonic and fetal mesenchymal stem cells. *Epigenetics* **6:** 52–62.

Fu Y, Rusznak Z, Herculano-Houzel S, Watson C, Paxinos G. 2012. Cellular composition characterizing postnatal development and maturation of the mouse brain and spinal cord. *Brain Struct Funct* doi: 10.1007/s00429-012-0462-x.

Garfield AS, Cowley M, Smith FM, Moorwood K, Stewart-Cox JE, Gilroy K, Baker S, Xia J, Dalley JW, Hurst LD, et al. 2011. Distinct physiological and behavioural functions for parental alleles of imprinted Grb10. *Nature* **469:** 534–538.

Gaszner M, Felsenfeld G. 2006. Insulators: Exploiting transcriptional and epigenetic mechanisms. *Nat Rev Genet* **7:** 703–713.

Grant CE, Bailey TL, Noble WS. 2011. FIMO: Scanning for occurrences of a given motif. *Bioinformatics* **27:** 1017–1018.

Gregg C, Zhang J, Butler J, Haig D, Dulac C. 2010a. Sex-specific parent-of-origin allelic expression in the mouse brain. *Science* **329:** 682–685.

Gregg C, Zhang J, Weissbourd B, Luo S, Schroth G, Haig D, Dulac C. 2010b. High-resolution analysis of parent-of-origin allelic expression in the mouse brain. *Science* **329:** 643–648.

Hadjur S, Williams L, Ryan N, Cobb B, Sexton T, Fraser P, Fisher A, Merkenschlager M. 2009. Cohesins form chromosomal *cis*-interactions at the developmentally regulated IFNG locus. *Nature* **460:** 410–413.

Hagiwara Y, Hirai M, Nishiyama K, Kanazawa I, Ueda T, Sakaki Y, Ito T. 1997. Screening for imprinted genes by allelic message display: Identification of a paternally expressed gene impact on mouse chromosome 18. *Proc Natl Acad Sci* **94:** 9249–9254.

Handoko L, Xu H, Li G, Ngan CY, Chew E, Schnapp M, Lee CW, Ye C, Ping JL, Mulawadi F, et al. 2011. CTCF-mediated functional chromatin interactome in pluripotent cells. *Nat Genet* **43:** 630–638.

Hark A, Schoenherr C, Katz D, Ingram R, Levorse J, Tilghman S. 2000. CTCF mediates methylation-sensitive enhancer-blocking activity at the H19/Igf2 locus. *Nature* **405:** 486–489.

Hikichi T, Kohda T, Kaneko-Ishino T, Ishino F. 2003. Imprinting regulation of the murine Meg1/Grb10 and human GRB10 genes; roles of brain-specific promoters and mouse-specific CTCF-binding sites. *Nucleic Acids Res* **31:** 1398–1406.

Kagey M, Newman J, Bilodeau S, Zhan Y, Orlando D, van Berkum N, Ebmeier C, Goossens J, Rahl P, Levine S, et al. 2010. Mediator and cohesin connect gene expression and chromatin architecture. *Nature* **467:** 430–435.

Kanduri C, Pant V, Loukinov D, Pugacheva E, Qi C-F, Wolffe A, Ohlsson R, Lobanenkov V. 2000. Functional association of CTCF with the insulator upstream of the *H19* gene is parent of origin-specific and methylation-sensitive. *Curr Biol* **10:** 853–856.

Keane TM, Goodstadt L, Danecek P, White MA, Wong K, Yalcin B, Heger A, Agam A, Slater G, Goodson M, et al. 2011. Mouse genomic variation and its effect on phenotypes and gene regulation. *Nature* **477:** 289–294.

Kim KP, Thurston A, Mummery C, Ward-van Oostwaard D, Priddle H, Allegrucci C, Denning C, Young L. 2007a. Gene-specific vulnerability to imprinting variability in human embryonic stem cell lines. *Genome Res* **17:** 1731–1742.

Kim TH, Abdullaev ZK, Smith AD, Ching KA, Loukinov DI, Green RD, Zhang MQ, Lobanenkov VV, Ren B. 2007b. Analysis of the vertebrate insulator protein CTCF-binding sites in the human genome. *Cell* **128:** 1231–1245.

Kim JK, Samaranayake M, Pradhan S. 2009. Epigenetic mechanisms in mammals. *Cell Mol Life Sci* **66:** 596–612.

Kobayashi H, Sakurai T, Sato S, Nakabayashi K, Hata K, Kono T. 2012. Imprinted DNA methylation reprogramming during early mouse embryogenesis at the *Gpr1–Zdbf2* locus is linked to long *cis*-intergenic transcription. *FEBS Lett* **586:** 827–833.

Kuzmin I, Geil L, Gibson L, Cavinato T, Loukinov D, Lobanenkov V, Lerman MI. 2005. Transcriptional regulator CTCF controls human interleukin 1 receptor-associated kinase 2 promoter. *J Mol Biol* **346:** 411–422.

Lee S, Wevrick R. 2000. Identification of novel imprinted transcripts in the Prader-Willi syndrome and Angelman syndrome deletion region: Further evidence for regional imprinting control. *Am J Hum Genet* **66:** 848–858.

Lefebvre L, Viville S, Barton S, Ishino F, Keverne E, Surani A. 1998. Abnormal maternal behaviour and growth retardation associated with loss of the imprinted gene *Mest*. *Nat Genet* **20:** 163–169.

Lewis A, Reik W. 2006. How imprinting centres work. *Cytogenet Genome Res* **113:** 81–89.

Li LC, Dahiya R. 2002. MethPrimer: Designing primers for methylation PCRs. *Bioinformatics* **18:** 1427–1431.

Li E, Beard C, Jaenisch R. 1993. Role for DNA methylation in genomic imprinting. *Nature* **366:** 362–365.

Li LL, Keverne EB, Aparicio SA, Ishino F, Barton SC, Surani MA. 1999. Regulation of maternal behavior and offspring growth by paternally expressed Peg3. *Science* **284:** 330–334.

Lin S, Ferguson-Smith A, Schultz R, Bartolomei M. 2011. Nonallelic transcriptional roles of CTCF and cohesins at imprinted loci. *Mol Cell Biol* **31:** 3094–3104.

Lossie AC, Whitney MM, Amidon D, Dong HJ, Chen P, Theriaque D, Hutson A, Nicholls RD, Zori RT, Williams CA, et al. 2001. Distinct phenotypes distinguish the molecular classes of Angelman syndrome. *J Med Genet* **38:** 834–845.

Lyck L, Kroigard T, Finsen B. 2007. Unbiased cell quantification reveals a continued increase in the number of neocortical neurones during early post-natal development in mice. *Eur J Neurosci* **26:** 1749–1764.

Manolio TA. 2010. Genomewide association studies and assessment of the risk of disease. *N Engl J Med* **363:** 166–176.

Martin D, Pantoja C, Fernandez Minan A, Valdes-Quezada C, Molto E, Matesanz F, Bogdanovic O, de la Calle-Mustienes E, Dominguez O, Taher L, et al. 2011. Genome-wide CTCF distribution in vertebrates defines equivalent sites that aid the identification of disease-associated genes. *Nat Struct Mol Biol* **18:** 708–714.

McCole R, Oakey R. 2008. Unwitting hosts fall victim to imprinting. *Epigenetics* **3:** 258–260.

Mukhopadhyay R, Yu W, Whitehead J, Xu J, Lezcano M, Pack S, Kanduri C, Kanduri M, Ginjala V, Vostrov A, et al. 2004. The binding sites for the chromatin insulator protein CTCF map to DNA methylation-free domains genome-wide. *Genome Res* **14:** 1594–1602.

Murrell A, Heeson S, Reik W. 2004. Interaction between differentially methylated regions partitions the imprinted genes *Igf2* and *H19* into parent-specific chromatin loops. *Nat Genet* **36:** 889–893.

Nativio R, Wendt K, Ito Y, Huddleston J, Uribe-Lewis S, Woodfine K, Krueger C, Reik W, Peters J-M, Murrell A. 2009. Cohesin is required for higher-order chromatin conformation at the imprinted *IGF2-H19* locus. *PLoS Genet* **5:** e1000739.

Neph S, Kuehn MS, Reynolds AP, Haugen E, Thurman RE, Johnson AK, Rynes E, Maurano MT, Vierstra J, Thomas S, et al. 2012. BEDOPS: High-performance genomic feature operations. *Bioinformatics* **28:** 1919–1920.

Nix D, Courdy S, Boucher K. 2008. Empirical methods for controlling false positives and estimating confidence in ChIP-seq peaks. *BMC Bioinformatics* **9:** 523.

Noonan JP, McCallion AS. 2010. Genomics of long-range regulatory elements. *Annu Rev Genomics Hum Genet* **11:** 1–23.

Ohlsson R. 2001. CTCF is a uniquely versatile transcription regulator linked to epigenetics and disease. *Trends Genet* **17:** 520–527.

Parelho V, Hadjur S, Spivakov M, Leleu M, Sauer S, Gregson H, Jarmuz A, Canzonetta C, Webster Z, Nesterova T, et al. 2008. Cohesins functionally associate with CTCF on mammalian chromosome arms. *Cell* **132:** 422–433.

Peters J, Williamson C. 2007. Control of imprinting at the Gnas cluster. *Epigenetics* **2:** 207–213.

Phillips JE, Corces VG. 2009. CTCF: Master weaver of the genome. *Cell* **137:** 1194–1211.

Plagge A, Isles A, Gordon E, Humby T, Dean W, Gritsch S, Fischer-Colbrie R, Wilkinson L, Kelsey G. 2005. Imprinted Nesp55 influences behavioral reactivity to novel environments. *Mol Cell Biol* **25:** 3019–3026.

Proudhon C, Duffie R, Ajjan S, Cowley M, Iranzo J, Carbajosa G, Saadeh H, Holland ML, Oakey RJ, Rakyan VK, et al. 2012. Protection against de novo methylation is instrumental in maintaining parent-of-origin methylation inherited from the gametes. *Mol Cell* **47:** 909–920.

Quenneville S, Verde G, Corsinotti A, Kapopoulou A, Jakobsson J, Offner S, Baglivo I, Pedone PV, Grimaldi G, Riccio A, et al. 2012. In embryonic stem cells, ZFP57/KAP1 recognize a methylated hexanucleotide to affect chromatin and DNA methylation of imprinting control regions. *Mol Cell* **44:** 361–372.

Reik W, Walter J. 2001. Genomic imprinting: Parental influence on the genome. *Nat Rev Genet* **2:** 21–32.

Rubio E, Reiss D, Welcsh P, Disteche C, Filippova G, Baliga N, Aebersold R, Ranish J, Krumm A. 2008. CTCF physically links cohesin to chromatin. *Proc Natl Acad Sci* **105:** 8309–8314.

Rugg-Gunn PJ, Ferguson-Smith AC, Pedersen RA. 2007. Status of genomic imprinting in human embryonic stem cells as revealed by a large cohort of independently derived and maintained lines. *Hum Mol Genet* **16:** R243–R251.

Sandhu KS, Shi C, Sjolinder M, Zhao Z, Gondor A, Liu L, Tiwari VK, Guibert S, Emilsson L, Imreh MP, et al. 2009. Nonallelic transvection of multiple imprinted loci is organized by the H19 imprinting control region during germline development. *Genes Dev* **23:** 2598–2603.

Schmidt D, Schwalie P, Ross-Innes C, Hurtado A, Brown G, Carroll J, Flicek P, Odom D. 2010. A CTCF-independent role for cohesin in tissue-specific transcription. *Genome Res* **20:** 578–588.

Schmidt D, Schwalie PC, Wilson MD, Ballester B, Goncalves A, Kutter C, Brown GD, Marshall A, Flicek P, Odom DT. 2012. Waves of retrotransposon expansion remodel genome organization and CTCF binding in multiple mammalian lineages. *Cell* **148:** 335–348.

Schulz R, Woodfine K, Menheniott TR, Bourc'his D, Bestor T, Oakey RJ. 2008. WAMIDEX: A web atlas of murine genomic imprinting and differential expression. *Epigenetics* **3:** 89–96.

Shen Y, Yue F, McCleary DF, Ye Z, Edsall L, Kuan S, Wagner U, Dixon J, Lee L, Lobanenkov VV, et al. 2012. A map of the *cis*-regulatory sequences in the mouse genome. *Nature* **488:** 116–120.

Shin H, Liu T, Manrai AK, Liu XS. 2009. CEAS: *Cis*-regulatory element annotation system. *Bioinformatics* **25:** 2605–2606.

Singh P, Wu X, Lee D-H, Li A, Rauch T, Pfeifer G, Mann J, Szabo P. 2011. Chromosome-wide analysis of parental allele-specific chromatin and DNA methylation. *Mol Cell Biol* **31:** 1757–1770.

Smallwood SA, Tomizawa S, Krueger F, Ruf N, Carli N, Segonds-Pichon A, Sato S, Hata K, Andrews SR, Kelsey G. 2011. Dynamic CpG island methylation landscape in oocytes and preimplantation embryos. *Nat Genet* **43:** 811–814.

Splinter E, Heath H, Kooren J, Palstra RJ, Klous P, Grosveld F, Galjart N, de Laat W. 2006. CTCF mediates long-range chromatin looping and local histone modification in the β-globin locus. *Genes Dev* **20:** 2349–2354.

Stedman W, Kang H, Lin S, Kissil J, Bartolomei M, Lieberman P. 2008. Cohesins localize with CTCF at the KSHV latency control region and at cellular c-*myc* and *H19/Igf2* insulators. *EMBO J* **27:** 654–666.

Szabo P, Pfeifer G, Mann J. 2004. Parent-of-origin-specific binding of nuclear hormone receptor complexes in the *H19-Igf2* imprinting control region. *Mol Cell Biol* **24:** 4858–4868.

Taatjes DJ. 2012. The human Mediator complex: A versatile, genome-wide regulator of transcription. *Trends Biochem Sci* **35:** 315–322.

Tsumura A, Hayakawa T, Kumaki Y, Takebayashi S, Sakaue M, Matsuoka C, Shimotohno K, Ishikawa F, Li E, Ueda HR, et al. 2006. Maintenance of self-renewal ability of mouse embryonic stem cells in the absence of DNA methyltransferases Dnmt1, Dnmt3a and Dnmt3b. *Genes Cells* **11:** 805–814.

Wendt K, Yoshida K, Itoh T, Bando M, Koch B, Schirghuber E, Tsutsumi S, Nagae G, Ishihara K, Mishiro T, et al. 2008. Cohesin mediates transcriptional insulation by CCCTC-binding factor. *Nature* **451:** 796–801.

Wilkins J. 2008. *Genomic imprinting*. Landes Bioscience/Springer, Austin, TX.

Williams C, Beaudet A, Clayton-Smith J, Knoll J, Kyllerman M, Laan L, Magenis E, Moncla A, Schinzel A, Summers J, et al. 2006. Angelman syndrome 2005: Updated consensus for diagnostic criteria. *Am J Med Genet* **140A:** 413–418.

Wood AJ, Roberts RG, Monk D, Moore GE, Schulz R, Oakey RJ. 2007. A screen for retrotransposed imprinted genes reveals an association between X chromosome homology and maternal germ-line methylation. *PLoS Genet* **3:** e20.

Wood AJ, Schulz R, Woodfine K, Koltowska K, Beechey CV, Peters J, Bourc'his D, Oakey RJ. 2008. Regulation of alternative polyadenylation by genomic imprinting. *Genes Dev* **22:** 1141–1146.

Xiao T, Wallace J, Felsenfeld G. 2011. Specific sites in the C terminus of CTCF interact with the SA2 subunit of the cohesin complex and are required for cohesin-dependent insulation activity. *Mol Cell Biol* **31:** 2174–2183.

Xie W, Barr CL, Kim A, Yue F, Young Lee A, Eubanks J, Dempster EL, Ren B. 2012. Base-resolution analyses of sequence and parent-of-origin dependent DNA methylation in the mouse genome. *Cell* **148:** 816–831.

Yalcin B, Wong K, Agam A, Goodson M, Keane TM, Gan X, Nellaker C, Goodstadt L, Nicod J, Bhomra A, et al. 2011. Sequence-based characterization of structural variation in the mouse genome. *Nature* **477:** 326–329.

Yoon B, Herman H, Hu B, Park Y, Lindroth A, Bell A, West A, Chang Y, Stablewski A, Piel J, et al. 2005. Rasgrf1 imprinting is regulated by a CTCF-dependent methylation-sensitive enhancer blocker. *Mol Cell Biol* **25:** 11184–11190.

Yusufzai TM, Felsenfeld G. 2004. The 5′-HS4 chicken β-globin insulator is a CTCF-dependent nuclear matrix-associated element. *Proc Natl Acad Sci* **101:** 8620–8624.

Received October 1, 2012; accepted in revised form June 20, 2013.

Appendix C

Custom Scripts

C.1 mapReadsToAlleles.pl

```
1  #!/apps/perl/5.15.1/bin/perl
2  #/usr/bin/perl
3
4  # Copyright 2010 Nikolaos Barkas
5  # All rights reserved.
6
7  # Author: Nikolaos Barkas
8  # Date: October 2010
9  # Description: A program for the assignment of new generation sequensing reads to alleles
10
11 # Libraries and pragmas
12 use strict;
13 use POSIX;
14 use Bio::DB::Sam;
15 use Getopt::Long;
16
17 # Input files - Change this to accept parameters from the command line
18 my $snpsFile;
19 my $bamFile;
20 my $showHelp;
21
22 &GetOptions(
23         "bam=s" => \$bamFile,
24         "snps=s" => \$snpsFile,
25         "help" => \$showHelp
26 );
27
28 if ($showHelp) {
29         usage();
30         exit 0;
31 }
32
33 if (!$bamFile || ! -e $bamFile) {
34         error("Bam file not found ". $bamFile);
35         usage();
36         exit;
37 }
38
39 if (!$snpsFile || ! -e $snpsFile ) {
40         error("Snps file not found ".$snpsFile);
41         usage();
42         exit;
43 }
44
45 sub usage {
46         print "mapReadsToAlleles --bam=bamFile --snp=snpsFile\n";
47 }
48
49 sub error {
50         print "Error: $_[0]\n";
51 }
52
53 # TODO: Load Genomic regions
54
55
56 # Load SNPs
57 # Initialise Chromosome SNP Hash
58 # Each element in the hash points to an array
59 # Which is an array of snps
60 my %snpChrHash = ();
61 open(SNPS, $snpsFile);
62 while (my $line = <SNPS>) {
63         chomp($line);
64         my @splitLine = split("\t",$line);
65         my $chr = lc($splitLine[0]);
```

```
66          push @{ $snpChrHash{$chr} },  \@splitLine  ;
67  }
68  close(SNPS);
69
70  # Sort the SNP Arrays by chromosomal position
71  # so that we can run binary search on these
72  foreach my $key (keys %snpChrHash) {
73          my @chrArray = @{$snpChrHash{$key}};
74          @chrArray = sort { @$a[1] <=> @$b[1] } @chrArray;
75  }
76
77  # Open Sam file
78  my $sam = Bio::DB::Sam->new( -bam => $bamFile);
79  my @alignments = $sam->features();
80
81  for my $a (@alignments) {
82          # A variable holding the output for each entry until the time
83          # it is to be printed to stdout. We write to this variable
84          # as opposed to stdout directy to allow for parallelisation of this loop
85          # across multiple cores without causing inconsistent output
86          my $lineBuffer = '';
87
88          # Information for this particular alignment
89          my $chromosome = $a->seq_id;
90          my $startRegion =  $a->start;
91          my $endRegion = $a->end;
92          my $primaryId = $a->primary_id;
93
94          # Print out the region information
95          $lineBuffer = $chromosome."\t".$startRegion."\t".$endRegion."\t";
96
97          # The array in the snp hash which holds snp info
98          # for the chromosome on which this read was mapped to
99          my @chromosomeSnps = @{$snpChrHash{lc($chromosome)}};
100
101         # Naive approach, loop over all snps
102         sub getSnpsInRegion {
103                 my @matchedSnps = ();
104                 my @chromosomeSnps = @{$_[0]};
105                 my $startRegion = $_[1];
106                 my $endRegion = $_[2];
107
108                 for my $i ( 0 .. $#{@chromosomeSnps} ) {
109                         my $snpPosition = $chromosomeSnps[$i][1];
110                         my $snpAlleleOne = $chromosomeSnps[$i][2];
111                         my $snpAlleleTwo = $chromosomeSnps[$i][3];
112
113                         if ( $startRegion <= $snpPosition && $snpPosition <= $endRegion) {
114                                 push( @matchedSnps, ['chromosome', $snpPosition, $snpAlleleOne,
                                        $snpAlleleTwo] );
115                         }
116                 }
117                 return @matchedSnps;
118         }
119
120         # Binary search into snp array
121         sub getSnpsInRegionBinary {
122                 my @chromosomeSnps = @{$_[0]};
123                 my $startRegion = $_[1];
124                 my $endRegion = $_[2];
125
126                 #Locate Start Position
127                 my $upperLimitPos = $#chromosomeSnps - 1;
128                 my $lowerLimitPos = 0;
129                 while ( $upperLimitPos > $lowerLimitPos + 1 ) {
130                         # Half-way through the existing range
131                         my $i = $lowerLimitPos + ceil(($upperLimitPos - $lowerLimitPos)/2);
132                         if ( $startRegion < $chromosomeSnps[$i][1] ) {
133                                 $upperLimitPos = $i;
134                         } elsif ($startRegion >= $chromosomeSnps[$i][1] ){
135                                 $lowerLimitPos = $i;
136                         }
137                 }
138                 my $regionStartIndex = $lowerLimitPos + 1;
139
140                 #Locate End Position
141                 my $upperLimitPos = $#chromosomeSnps -1;
142                 my $lowerLimitPos =  0; #$regionStartIndex; # Clearly the end position will be
                        after the start
143                 while ( $upperLimitPos > $lowerLimitPos + 1 ) {
144                         # Half-way through existing range
145                         my $i = $lowerLimitPos + ceil(($upperLimitPos - $lowerLimitPos)/2);
146                         if ( $endRegion < $chromosomeSnps[$i][1] ) {
147                                 $upperLimitPos = $i;
148                         } elsif ( $endRegion >= $chromosomeSnps[$i][1] ){
149                                 $lowerLimitPos = $i;
150                         }
151                 }
152                 my $regionEndIndex = $upperLimitPos - 1;
153
154                 # Get array slice with SNPs of interest
155                 my @g = @chromosomeSnps[$regionStartIndex .. $regionEndIndex];
156                 return @g;
157         }
158
159         # SNP lookup method
160         my @matchedSnps = getSnpsInRegionBinary(\@chromosomeSnps, $startRegion, $endRegion);
161
162         # Alternative linear search, much much slower
163         # my @matchedSnps = getSnpsInRegion(\@chromosomeSnps, $startRegion, $endRegion);
164
165         # From the SNPs we have just found in the  read in the region
166         # we need to pick the right one to use to do the assignment with
```

```
167
168             # The scores of the SNPs
169             my @scores = $a->qscore;
170
171             # The mapping quality of the read
172             my $match_qual = $a->qual;
173
174             # Read start position on the reference
175             my $readStart = $a->start;
176
177             #  Best hit counters
178             my $maxScoreValue = 0;
179             my $maxScorePosition = -1;
180             my $maxScoreSnpIndex = 0;
181
182             sub positionInBedRegion {
183                     return 1;
184             }
185
186             # Best SNP selection
187             for my $j ( 0 .. $#matchedSnps) {
188                     my $snpPosition =  $matchedSnps[$j][1];
189                     my $snpPositionInRead = $snpPosition - $readStart;
190
191                     # TODO: Apply extra filter to ensure that the actual
192                     # SNP used mapped to a genomic position which is in
193                     # a region on interest. This is particularly important
194                     # for RNA-seq data were systematic artifacts may arise
195                     # if flanking genomic regions are used.
196
197                     # Pick the best SNP
198                     if ($scores[$snpPositionInRead] > $maxScoreValue && positionInBedRegion(
                            $snpPosition)) {
199                             $maxScoreValue = @scores[$snpPositionInRead];
200                             $maxScorePosition = $snpPositionInRead;
201                             $maxScoreSnpIndex = $j;
202                     }
203             }
204
205             # Get the base read in the position corresponding to the snp
206             my $snpReadBase = substr($a->query->dna, $maxScorePosition, 1);
207
208             # Get the possible allele bases
209             my $alleleOneBase = $matchedSnps[$maxScoreSnpIndex][2];
210             my $alleleTwoBase = $matchedSnps[$maxScoreSnpIndex][3];
211
212             # Do assignment
213             # Base one is alwats ATGC in the input file
214             if ($snpReadBase eq $alleleOneBase ) {
215                     $lineBuffer = $lineBuffer .  '1';
216             } else {
217                     # But base 2 may use abbreviations
218                     # The abbreviations are ('http://www.chem.qmul.ac.uk/iubmb/misc/naseq.html#201')
219                     # R --> AG
220                     # Y --> CT
221                     # W --> AT
222                     # S --> GC
223                     # M --> AC
224                     # K --> GT
225                     # H --> !G
226                     # B --> !A
227                     # V --> !T
228                     # D --> !C
229                     # N --> ATGC
230                     # Of these only the pairs are used in the input
231
232                     if ( $snpReadBase eq $alleleTwoBase ) {
233                             # The simple case
234                             $lineBuffer = $lineBuffer . '2';
235                     } elsif ( $alleleTwoBase eq "R" ) {
236                             if ( $snpReadBase eq "A" || $snpReadBase eq "G" ) {
237                                     $lineBuffer = $lineBuffer . '2';
238                             } else {
239                                     $lineBuffer = $lineBuffer . 'X';
240                             }
241                     } elsif ( $alleleTwoBase eq "Y" ) {
242                             if ( $snpReadBase eq "C" || $snpReadBase eq "T" ) {
243                                     $lineBuffer = $lineBuffer . '2';
244                             } else {
245                                     $lineBuffer = $lineBuffer . 'X';
246                             }
247                     } elsif ( $alleleTwoBase eq "W" ) {
248                             if ( $snpReadBase eq "A" || $snpReadBase eq "T" ) {
249                                     $lineBuffer = $lineBuffer . '2';
250                             } else {
251                                     $lineBuffer = $lineBuffer . 'X';
252                             }
253                     } elsif ( $alleleTwoBase eq "S" ) {
254                             if ( $snpReadBase eq "G" || $snpReadBase eq "C" ) {
255                                     $lineBuffer = $lineBuffer . '2';
256                             } else {
257                                     $lineBuffer = $lineBuffer . 'X';
258                             }
259                     } elsif ( $alleleTwoBase eq "M" ) {
260                             if ( $snpReadBase eq "A" || $snpReadBase eq "C" ) {
261                                     $lineBuffer = $lineBuffer . '2';
262                             } else {
263                                     $lineBuffer = $lineBuffer . 'X';
264                             }
265                     } elsif ( $alleleTwoBase eq "K" ) {
266                             if ( $snpReadBase eq "G" || $snpReadBase eq "T" ) {
267                                     $lineBuffer = $lineBuffer . '2';
268                             } else {
```

```
269                                     $lineBuffer = $lineBuffer . 'X';
270                                 }
271                         } else {
272                                 # Unknown Allele 2 code
273                                 $lineBuffer = $lineBuffer . 'X';
274                         }
275                 }
276                 $lineBuffer = $lineBuffer . "\t" .$maxScoreValue . "\t" . $match_qual . "\t" .
                        $primaryId;
277
278                 # TODO: Substitute this with a thread safe version and parallelise
279                 printf($lineBuffer."\n");
280         }
```

C.2 methExtractorToBasepair.pl

```
1
2   #!/usr/bin/perl
3
4   use strict;
5
6   my %methStatus = ();
7
8   while(<>) {
9           chomp();
10          my @fields = split /\t/;
11
12          my $chr = $fields[2];
13          my $pos = $fields[3];
14          my $mStatus = $fields[1];
15
16          $methStatus{$chr}{$pos}{$mStatus}++;
17  }
18
19  foreach my $key (sort keys %methStatus) {
20          foreach my $key2 ( sort keys %{ $methStatus{$key} } ) {
21                  my $me;
22                  my $unme;
23
24                  if (defined $methStatus{$key}{$key2}{"+"}) {
25                          $me = $methStatus{$key}{$key2}{"+"};
26                  } else {
27                          $me = 0;
28                  }
29
30                  if (defined $methStatus{$key}{$key2}{"-"}) {
31                          $unme = $methStatus{$key}{$key2}{"-"};
32                  } else {
33                          $unme = 0;
34                  }
35
36                  my $count = $me + $unme;
37                  if ($count > 0) {
38                          my $percent = $me / $count * 100;
39                          print "$key\t$key2\t$percent\t$count\n";
40                  }
41          }
42  }
```

C.3 compMethExpr.tex

```
1
2   library("GenomicRanges")
3   library("ggplot2")
4
5   # Load CGIs
6   cgi_meth <- read.csv('input/results.csv')
7   mean.meth <- apply(X=cgi_meth[,c("ec_rep1","ec_rep2","et_rep1","et_rep2")],MARGIN=1 , FUN=mean)
8   cgi_meth.gr <- GRanges(seqnames = Rle(cgi_meth$chr),
9                          ranges = IRanges(start=cgi_meth$start, end=cgi_meth$end))
10  cgi_meth.gr$methylation <- mean.meth
11  rm(cgi_meth, mean.meth)
12
13  # Load gene expression
14  gene_exp <- read.table('input/gene_exp.diff',header=T,as.is=T)
15  seqnames <- factor(sapply(strsplit(gene_exp[["locus"]],':'),'[',1))
16  positions <- sapply(strsplit(gene_exp[["locus"]],':'),'[',2)
17  start <- as.numeric(sapply(strsplit(positions,'-'),'[',1))
18  end <- as.numeric(sapply(strsplit(positions,'-'),'[',2))
19  expr <- apply(gene_exp[,c("value_1","value_2")], MARGIN=1, mean)
20  gene_exp.gr <- GRanges(seqnames= Rle(seqnames), ranges = IRanges(start=start, end=start+1))
21  gene_exp.gr$expr <-  expr
22  rm(expr, start, end, positions, seqnames, gene_exp)
23
24
25  gaps <- data.frame(gap=c(seq(0,10000,100),seq(10000,300000,1000)))
26
27  getOverlapSignificance <- function(x) {
28    hits <- findOverlaps(cgi_meth.gr, gene_exp.gr,maxgap = x)
29    comp <- data.frame(methylation=cgi_meth.gr[queryHits(hits),]$methylation,expr=gene_exp.gr[
          subjectHits(hits),]$expr)
30    comp$meth.level <- cut(comp$methylation, breaks=c(-1,50,100), labels = c("low","high"))
31    high <- subset(comp, meth.level =="high")
```

```
32    low <- subset(comp, meth.level =="low")
33    t.test(x=high$expr,y=low$expr)$p.value
34  }
35
36  getOverlapRatio <- function(x) {
37    hits <- findOverlaps(cgi_meth.gr, gene_exp.gr,maxgap = x)
38    comp <- data.frame(methylation=cgi_meth.gr[queryHits(hits),]$methylation,expr=gene_exp.gr[
          subjectHits(hits),]$expr)
39    comp$meth.level <- cut(comp$methylation, breaks=c(-1,50,100), labels = c("low","high"))
40    high <- subset(comp, meth.level =="high")
41    low <- subset(comp, meth.level =="low")
42    mean(high$expr+0.01)/mean(low$expr+0.01)
43  }
44
45  p.vals <- apply(gaps, MARGIN=1, getOverlapSignificance)
46  ratios <- apply(gaps, MARGIN=1, getOverlapRatio)
47
48  library('scales')
49
50  library('reshape2')
51  res <- data.frame(distance=gaps$gap, p.value=p.vals,ratio=ratios,significant=(p.vals < 0.05))
52  res.m <- melt(res,id.vars = c("distance","significant"))
53  p <- ggplot(res.m, aes(x=distance,y=value,color=significant)) + geom_point() +
54    facet_grid(variable~.,scales="free_y") + theme_bw() + scale_y_continuous(name="") +
55    scale_x_continuous(labels=comma,name="Distance_in_basepairs")
56  ggsave('output/methExprDista.png')
57  p <- ggplot(subset(res.m, variable=="ratio"), aes(x=distance,y=value,color=significant)) +
        geom_point() + geom_smooth() +
58     theme_bw() + scale_y_continuous(name="") +
59    scale_x_continuous(labels=comma,name="Distance_in_basepairs",limit=c(0,10000))
60  ggsave('output/methExprDistaZoom.png')
61
62  # And a boxplot for immediate overlaps
63
64  hits <- findOverlaps(cgi_meth.gr, gene_exp.gr)
65  comp <- data.frame(methylation=cgi_meth.gr[queryHits(hits),]$methylation,expr=gene_exp.gr[
        subjectHits(hits),]$expr)
66  comp$meth.level <- cut(comp$methylation, breaks=c(-1,50,100), labels = c("low","high"))
67  high <- subset(comp, meth.level =="high")
68  low <- subset(comp, meth.level =="low")
69
70  p <- ggplot(comp,aes(x=meth.level,y=log(expr))) + geom_boxplot() + theme_bw() +
71    scale_x_discrete(name="Methylation_Level") + scale_y_continuous(name="log(Mean_Expression)")
72  ggsave('output/boxPlotImmediate.png')
73
74  getOverlapSignificance(0)
75
76  ##########
77  # Now do the same thing per sample (EC)
78  ##########
79  # Load CGIs
80  cgi_meth <- read.csv('input/results.csv')
81  mean.meth <- apply(X=cgi_meth[,c("ec_rep1","ec_rep2")],MARGIN=1 , FUN=mean)
82  cgi_meth.gr <- GRanges(seqnames = Rle(cgi_meth$chr),
83                         ranges = IRanges(start=cgi_meth$start, end=cgi_meth$end))
84  cgi_meth.gr$methylation <- mean.meth
85  rm(cgi_meth, mean.meth)
86
87  # Load gene expression
88  gene_exp <- read.table('input/gene_exp.diff',header=T,as.is=T)
89  seqnames <- factor(sapply(strsplit(gene_exp[["locus"]],':'),'[',1))
90  positions <- sapply(strsplit(gene_exp[["locus"]],':'),'[',2)
91  start <- as.numeric(sapply(strsplit(positions,'-'),'[',1))
92  end <- as.numeric(sapply(strsplit(positions,'-'),'[',2))
93  expr <- gene_exp[,c("value_1")]
94  gene_exp.gr <- GRanges(seqnames= Rle(seqnames), ranges = IRanges(start=start, end=start+1))
95  gene_exp.gr$expr <-  expr
96  rm(expr, start, end, positions, seqnames, gene_exp)
97
98  p.vals <- apply(gaps, MARGIN=1, getOverlapSignificance)
99  ratios <- apply(gaps, MARGIN=1, getOverlapRatio)
100
101 res <- data.frame(distance=gaps$gap, p.value=p.vals,ratio=ratios,significant=(p.vals < 0.05))
102 res.m <- melt(res,id.vars = c("distance","significant"))
103 p <- ggplot(res.m, aes(x=distance,y=value,color=significant)) + geom_point() +
104   facet_grid(variable~.,scales="free_y") + theme_bw() + scale_y_continuous(name="") +
105   scale_x_continuous(labels=comma,name="Distance_in_basepairs")
106 ggsave('output/ec.methExprDista.png')
107 p <- ggplot(subset(res.m, variable=="ratio"), aes(x=distance,y=value,color=significant)) +
        geom_point() + geom_smooth() +
108   theme_bw() + scale_y_continuous(name="") +
109   scale_x_continuous(labels=comma,name="Distance_in_basepairs",limit=c(0,10000))
110 ggsave('output/ec.methExprDistaZoom.png')
111
112 # And a boxplot for immediate overlaps
113
114 hits <- findOverlaps(cgi_meth.gr, gene_exp.gr)
115 comp <- data.frame(methylation=cgi_meth.gr[queryHits(hits),]$methylation,expr=gene_exp.gr[
        subjectHits(hits),]$expr)
116 comp$meth.level <- cut(comp$methylation, breaks=c(-1,50,100), labels = c("low","high"))
117 high <- subset(comp, meth.level =="high")
118 low <- subset(comp, meth.level =="low")
119
120 p <- ggplot(comp,aes(x=meth.level,y=log(expr))) + geom_boxplot() + theme_bw() +
121   scale_x_discrete(name="Methylation_Level") + scale_y_continuous(name="log(Mean_Expression)")
122 ggsave('output/ec.boxPlotImmediate.png')
123
124 getOverlapSignificance(0)
125
126 ##########
127 #  do the same thing per sample (ET)
128 ##########
129 # Load CGIs
```

```
cgi_meth <- read.csv('input/results.csv')
mean.meth <- apply(X=cgi_meth[,c("et_rep1","et_rep2")],MARGIN=1 , FUN=mean)
cgi_meth.gr <- GRanges(seqnames = Rle(cgi_meth$chr),
                       ranges = IRanges(start=cgi_meth$start, end=cgi_meth$end))
cgi_meth.gr$methylation <- mean.meth
rm(cgi_meth, mean.meth)

# Load gene expression
gene_exp <- read.table('input/gene_exp.diff',header=T,as.is=T)
seqnames <- factor(sapply(strsplit(gene_exp[["locus"]],':'),'[',1))
positions <- sapply(strsplit(gene_exp[["locus"]],':'),'[',2)
start <- as.numeric(sapply(strsplit(positions,'-'),'[',1))
end <- as.numeric(sapply(strsplit(positions,'-'),'[',2))
expr <- gene_exp[,c("value_2")]
gene_exp.gr <- GRanges(seqnames= Rle(seqnames), ranges = IRanges(start=start, end=start+1))
gene_exp.gr$expr <-  expr
rm(expr, start, end, positions, seqnames, gene_exp)

p.vals <- apply(gaps, MARGIN=1, getOverlapSignificance)
ratios <- apply(gaps, MARGIN=1, getOverlapRatio)

res <- data.frame(distance=gaps$gap, p.value=p.vals,ratio=ratios,significant=(p.vals < 0.05))
res.m <- melt(res,id.vars = c("distance","significant"))
p <- ggplot(res.m, aes(x=distance,y=value,color=significant)) + geom_point() +
  facet_grid(variable~.,scales="free_y") + theme_bw() + scale_y_continuous(name="") +
  scale_x_continuous(labels=comma,name="Distance_in_basepairs")
ggsave('output/et.methExprDista.png')
p <- ggplot(subset(res.m, variable=="ratio"), aes(x=distance,y=value,color=significant)) +
    geom_point() + geom_smooth() +
  theme_bw() + scale_y_continuous(name="") +
  scale_x_continuous(labels=comma,name="Distance_in_basepairs",limit=c(0,10000))
ggsave('output/et.methExprDistaZoom.png')

#  a boxplot for immediate overlaps
hits <- findOverlaps(cgi_meth.gr, gene_exp.gr)
comp <- data.frame(methylation=cgi_meth.gr[queryHits(hits),]$methylation,expr=gene_exp.gr[
    subjectHits(hits),]$expr)
comp$meth.level <- cut(comp$methylation, breaks=c(-1,50,100), labels = c("low","high"))
high <- subset(comp, meth.level =="high")
low <- subset(comp, meth.level =="low")


p <- ggplot(comp,aes(x=meth.level,y=log(expr))) + geom_boxplot() + theme_bw() +
  scale_x_discrete(name="Methylation_Level") + scale_y_continuous(name="log(Mean_Expression)")
ggsave('output/et.boxPlotImmediate.png')

getOverlapSignificance(0)

save.image('final.RData')
```

C.4 getTFsByGOterm.pl

```
1
2  drop table if exists tfs;
3
4
5  create table tfs select gene_product.symbol from
6          mygo.species,
7          mygo.gene_product,
8          mygo.association,
9          mygo.term
10 where
11         mygo.species.genus="Mus" and
12         mygo.species.species="musculus" and
13         mygo.gene_product.species_id=mygo.species.id and
14         mygo.association.is_not = 0 and
15         mygo.association.term_id = mygo.term.id and
16         mygo.association.gene_product_id = mygo.gene_product.id and
17         mygo.term.acc="GO:0003700";
18
19
20 drop table if exists up;
21 drop table if exists down;
22
23 create table up(gene varchar(255));
24 create table down(gene varchar(255));
25
26 load data local infile 'tmp/Upregulated.txt' into table up;
27 load data local infile 'tmp/Downregulated.txt' into table down;
28
29
30 select distinct tfs.symbol from tfs, up where tfs.symbol = up.gene;
31 select distinct tfs.symbol from tfs, down where tfs.symbol = down.gene;
```

C.5 encodeTFdistribution.R

```
1  library("RMySQL")
2  library("GenomicRanges")
3  library("GenomicFeatures")
4  library('ChIPseeker')
5  library('TxDb.Mmusculus.UCSC.mm9.knownGene')
6  library('clusterProfiler')
7
```

```
con = dbConnect(RMySQL::MySQL(), host = "genome-mysql.cse.ucsc.edu", user = "genome", password =
      "", dbname = "mm9")

plotEncodeTSSdistribution <- function(con=NULL,tableName=NULL,txdb=TxDb.Mmusculus.UCSC.mm9.
    knownGene, searchRadius=10000)
{
  query <- paste0("select * from ",tableName)
  res <-  dbGetQuery(con,query)
  res$strand[res$strand == '.']="*"
  peaks <- GRanges(seqnames=Rle(res$chrom),ranges=IRanges(res$chromStart, end=res$chromEnd,
      names=res$name), strand=Rle(res$strand),score=res$score)
  promoter <- getPromoters(TxDb=txdb, upstream=searchRadius, downstream=searchRadius)
  tagMatrix <- getTagMatrix(peaks, windows=promoter)
  plot(plotAvgProf(tagMatrix, xlim=c(-searchRadius, searchRadius), xlab="Genomic Region (5'->3')
      ",y="Peak Density"))
}

makePlot <- function(tableName=NULL) {
  pdf(paste0('output/',tableName,'.pdf'))
  plotEncodeTSSdistribution(con=con,tableName=tableName,searchRadius=20000)
  dev.off()
}

tables <- read.csv('tables.csv',as.is=T)

for (i in 1:dim(tables)[1]) {
  makePlot(tables[i,])
}
```

C.6 getTSSsequences.R

```
library('org.Mm.eg.db')
library("GenomicRanges")
library("TxDb.Mmusculus.UCSC.mm9.knownGene")
library("BSgenome.Mmusculus.UCSC.mm9")
library("BSgenome")
library("Biostrings")

stable.genes.symbol <- read.csv('genesForMotif//stable_genes.csv',as.is=T,header=F)
upregulated.genes.symbol <- read.csv('genesForMotif//upregulated.txt',as.is=T,header=F)

db <- TxDb.Mmusculus.UCSC.mm9.knownGene
all.promoters.mm9 <- promoters(db,upstream=5000,downstream=5000)
kgXref <- read.csv('kgXref.mm9.csv')
ucsc2symbol <- kgXref[,c("X.kgID","geneSymbol")]

stable.genes.ucsc <- ucsc2symbol[which(ucsc2symbol$geneSymbol %in% stable.genes.symbol[,1]),]
up.genes.ucsc <- ucsc2symbol[which(ucsc2symbol$geneSymbol %in% upregulated.genes.symbol[,1]),]

stable.promoters <- all.promoters.mm9[which(all.promoters.mm9$tx_name %in% stable.genes.ucsc$X.
    kgID),]
up.promoters <- all.promoters.mm9[which(all.promoters.mm9$tx_name %in% up.genes.ucsc$X.kgID),]

makeDirectionalPromoterWindow <- function(promoters, startOffset, windowSize) {
  windowSize <- windowSize -1

  promoters.pos <- promoters[strand(promoters) == '+']
  promoters.neg <- promoters[strand(promoters) == '-']

  start(ranges(promoters.pos)) <- start(ranges(promoters.pos)) + startOffset
  end(ranges(promoters.pos)) <- start(ranges(promoters.pos)) + windowSize

  end(ranges(promoters.neg)) <- end(ranges(promoters.neg)) - startOffset
  start(ranges(promoters.neg)) <- end(ranges(promoters.neg)) - windowSize

  c(promoters.pos,promoters.neg)
}

writeGenomicRangesAsBed <- function(gr,bedFile) {
        df <- data.frame(seqnames=seqnames(gr),
                  starts=start(gr)-1,
                  ends=end(gr),
                  names=gr$tx_name,
                  scores=c(rep(".", length(gr))),
                  strands=strand(gr))

        write.table(df, file=bedFile, quote=F, sep="\t", row.names=F, col.names=F)
}

mm9 <- BSgenome.Mmusculus.UCSC.mm9

windowSize <- 1000
windowStep <- 500
for (i in seq(0,9000,windowStep)) {
  stable.promoters.window <- makeDirectionalPromoterWindow(stable.promoters,i,windowSize)

  # Make the windows unique across the genome
  stable.promoters.window <- stable.promoters.window[!duplicated(stable.promoters.window)]


  writeGenomicRangesAsBed(stable.promoters.window,paste0("output/stable_window_",i,".bed"))
  stable.seq <- getSeq(x=mm9,names=seqnames(stable.promoters.window),
          start=start(stable.promoters.window),
          end=end(stable.promoters.window),
          strand=strand(stable.promoters.window))
  names(stable.seq) <- stable.promoters.window$tx_name
  writeXStringSet(stable.seq,paste0("output/stable_window_",i,".fasta"))

```

```
67
68      up.promoters.window <- makeDirectionalPromoterWindow(up.promoters,i,windowSize)
69      up.promoters.window <- up.promoters.window[!duplicated(up.promoters.window)]
70
71      writeGenomicRangesAsBed(up.promoters.window,paste0("output/up_window_",i,".bed"))
72      up.seq <- getSeq(x=mm9,names=seqnames(up.promoters.window),
73                               start=start(up.promoters.window),
74                               end=end(up.promoters.window),
75                               strand=strand(up.promoters.window))
76      names(up.seq) <- up.promoters.window$tx_name
77      writeXStringSet(up.seq,paste0("output/up_window_",i,".fasta"))
78  }
```